|  | **Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное бюджетное образовательное учреждение**  **высшего образования**  **«Московский государственный технический университет**  **имени Н.Э. Баумана**  **(национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |
| --- | --- |

ФАКУЛЬТЕТ «Информатика, искусственный интеллект и системы управления»

КАФЕДРА «Системы обработки информации и управления»

**РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

***К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ***

***НА ТЕМУ:***

***Система оперативного анализа данных***

***«Анализ результатов олимпийских игр с 1896 года»***

Студент ИУ5-55Б **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** Т.Н. Киндыкбаев

(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** К.Ю. Маслеников

(Подпись, дата) (И.О.Фамилия)

Консультант **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

(Подпись, дата) (И.О.Фамилия)

*2022 г.*

**Министерство науки и высшего образования Российской Федерации**

**Федеральное государственное бюджетное образовательное учреждение**

**высшего образования**

**«Московский государственный технический университет имени Н.Э. Баумана**

**(национальный исследовательский университет)»**

**(МГТУ им. Н.Э. Баумана)**

УТВЕРЖДАЮ

Заведующий кафедрой \_\_\_\_\_\_\_\_\_\_

(Индекс)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(И.О.Фамилия)

« \_\_\_\_\_ » \_\_\_\_\_\_\_\_\_\_\_\_ 20 \_\_\_\_ г.

**ЗАДАНИЕ**

**на выполнение научно-исследовательской работы**

по теме \_\_\_\_ Анализ данных по базе данных «Анализ результатов олимпийских игр с 1896 года»\_\_\_\_\_\_\_\_\_\_\_\_

Студент группы \_\_\_\_\_\_\_ИУ5-55Б\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Киндыкбаев Тимур Нуржигитович\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Фамилия, имя, отчество)

Направленность НИР (учебная, исследовательская, практическая, производственная, др.)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Производственная\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Источник тематики (кафедра, предприятие, НИР) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_НИР\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

График выполнения НИР: 25% к 3 нед., 50% к 9 нед., 75% к 12 нед., 100% к 15 нед.

***Техническое задание***  *Спроектировать систему анализа и визуализации по теме «Продажи видеоигр», провести визуализацию полученных данных посредством языка программирования Python*

***Оформление научно-исследовательской работы:***

Расчетно-пояснительная записка на 20 листах формата А4.

Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Дата выдачи задания «10» сентября 2022 г.

**Руководитель НИР**  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ К.Ю. Маслеников

(Подпись, дата) (И.О.Фамилия)

**Студент \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** Т.Н.Киндыкбаев

(Подпись, дата) (И.О.Фамилия)

Примечание: Задание оформляется в двух экземплярах: один выдается студенту, второй хранится на кафедре.

**СОДЕРЖАНИЕ**

[***ВВЕДЕНИЕ 4***](#_heading=h.gjdgxs)

[***1***](#_heading=h.3znysh7) ***Определение данных для анализа 5***

[***1***](#_heading=h.3dy6vkm) ***Формулирование гипотез 6***

[***2***](#_heading=h.1t3h5sf) ***Общее исследование данных 7***

[**3.1**](#_heading=h.4d34og8) **Получение информации о данных 8**

[**3.3**](#_heading=h.2s8eyo1) **Проверка названия столбцов 9**

[***3***](#_heading=h.17dp8vu) ***Исследовательский анализ 11***

[**4.1**](#_heading=h.3rdcrjn) **Страны, отправившие наибольшее количество участников 11**

[**4.2**](#_heading=h.26in1rg) **Участники Олимпийских игр, нанесенные на карту 12**

[**4.3**](#_heading=h.lnxbz9) **Анализ результатов медального зачета 13**

[**4.4**](#_heading=h.35nkun2) **Матрица зависимости переменных 17**

[**4.5**](#_heading=h.1ksv4uv) **Определение и анализ эффективности выступления 18**

[***4***](#_heading=h.44sinio) ***Проверка гипотез 20***

[***5***](#_heading=h.2jxsxqh) ***Общий вывод 23***

[***ЗАКЛЮЧЕНИЕ 24***](#_heading=h.z337ya)

[***СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 25***](#_heading=h.3j2qqm3)

# ВВЕДЕНИЕ

Цель:

Исследовать базу данных и выявить закономерности в результатах Олимпийских игр, чтобы предположить результаты медального зачета на следующий год.

Задачи:

1. Определение данных.
2. Формулирование гипотез.
3. Загрузка данных в Python.
4. Проверка данных.
5. Очистка данных.
6. Преобразование данных.
7. Выбор данных для анализа.
8. Агрегирование данных.
9. Визуализация данных.
10. Подтверждение или опровержение поставленных гипотез.
11. Формулирование ограничений и выводов.

# Определение данных для анализа

В качестве данных для анализа был выбран dataset «Результаты Олимпийских игр с 1896 г».

В наборе данных представлены суммарное количество золотых, серебряных и бронзовых медалей за всю историю Олимпийских игр для каждой страны.

Данные содержат следующие столбцы:

1. сountries - название страны
2. ioc\_code - код страны
3. summer\_participations - количество участников от страны на летних Олимпийских играх
4. summer\_gold – количество золотых медалей на летних Олимпийских играх
5. summer\_silver - количество серебряных медалей на летних Олимпийских играх
6. summer\_bronze - количество бронзовых медалей на летних Олимпийских играх
7. summer\_total - суммарное количество медалей на летних Олимпийских играх
8. winter\_participations - количество участников от страны на зимних Олимпийских играх
9. winter\_gold - количество золотых медалей на зимних Олимпийских играх
10. winter\_silve - количество серебряных медалей на зимних Олимпийских играх
11. winter\_bronze - количество бронзовых медалей на зимних Олимпийских играх
12. winter\_total - суммарное количество медалей на зимних Олимпийских играх
13. total\_participation – суммарное количество участников на обоих видах Олимпийских игр
14. total\_gold – суммарное количество золотых медалей на обоих видах Олимпийских игр
15. total\_silver - суммарное количество серебряных медалей на обоих видах Олимпийских игр
16. total\_bronze - суммарное количество бронзовых медалей на обоих видах Олимпийских игр

# Формулирование гипотез

Гипотеза 1:

*Нулевая гипотеза H0*: Распределение результатов медального зачета нормально

*Альтернативная гипотеза H1:* Распределение результатов медального зачета ненормально

Гипотеза 2:

*Нулевая гипотеза:* Средние результаты, показываемые Северной Европой на зимних Олимпийских играх, не отличаются от результатов остальных странах

*Альтернативна гипотеза:* Средние результаты, показываемые Северной Европой на зимних Олимпийских играх, выше, чем в остальных странах

# Общее исследование данных

Анализ будет проводиться с помощью средств Python, используются библиотеки: pandas, numpy, matplotlib, seaborn и scipy. Python – наиболее удобная и быстрая среда для анализа данных. Удостоверимся, что все библиотеки загружены и подключены. Загрузим данные из csv файла в DataFrame.

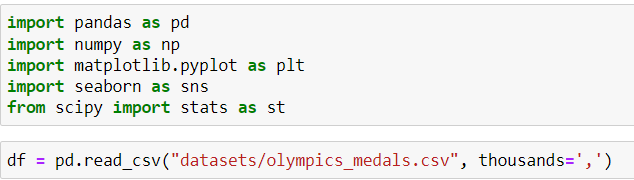
****

Рисунок 1 – Загрузка данных

Удостоверимся, что данные загружены нормально и посмотрим на получившуюся таблицу, проанализируем первые 10 строк, удостоверимся, что числа загрузились нормально – максимальные занчение достигнуты и не пропущены.

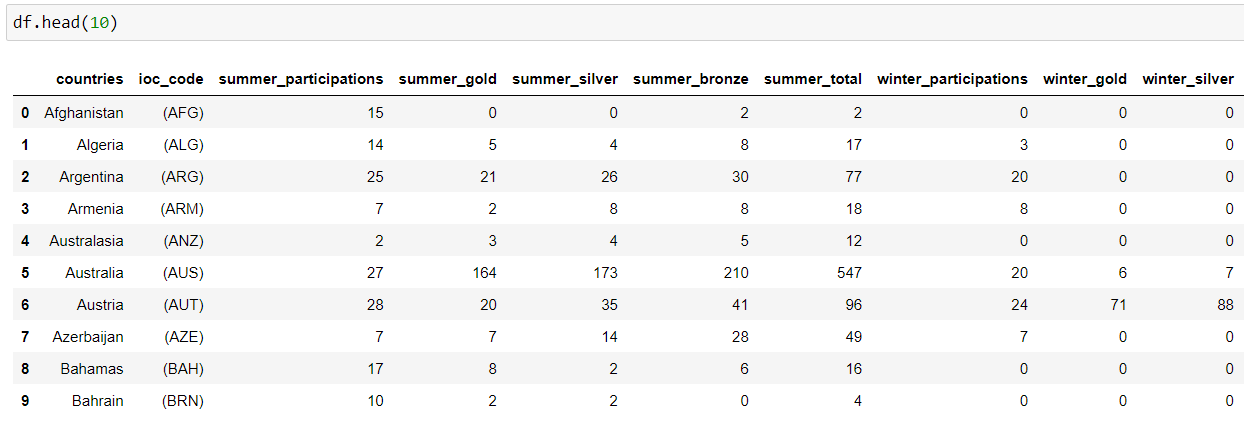


Рисунок 2 - Изображение таблицы

**3.1 Получение информации о данных**

Исследуем общую информацию о данных. Посмотрим, какого типа присутствуют данные в таблице, их количество и сколько памяти занимает таблица. Рассмотрим это с помощью метода data.info()[5].

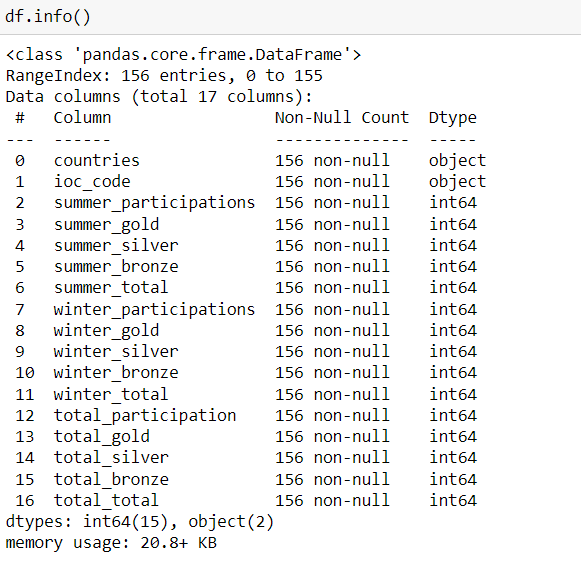


Рисунок 3 - Информация о данных

Заметим, что пропусков в данных нет из-за того, что количество строк в каждом столбце и максимальное количество строк везде совпадают и удостоверимся в этом еще раз:

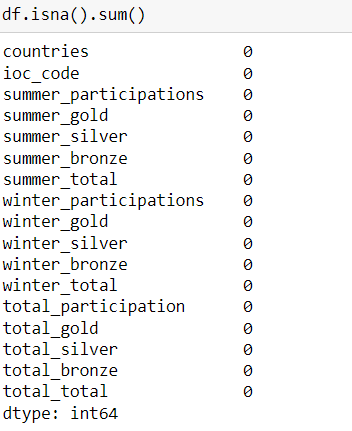
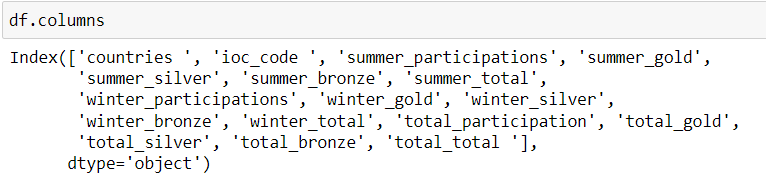
****

Рисунок 4 - Проверка данных

В данном случае пропуски заполнять не нужно, все данные заполнены, информация полная.

Проверим на звания столбцов

## 3.3 Проверка названия столбцов

****

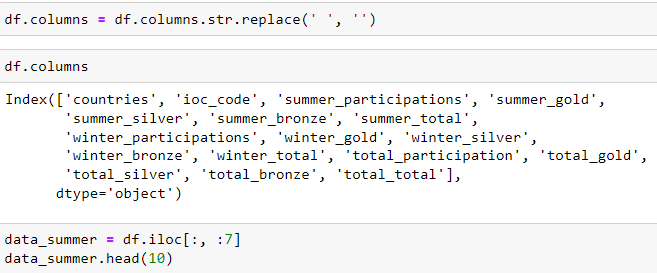


Рисунок 5 - Проверка столбцов

Посчитаем количество дубликатов, используя известную команду:

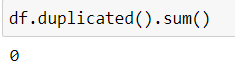
****

Рисунок 6 - Проверка на дубликаты

Отформатируем значение столбца ioc\_code, так как данные представлены в неудобном формате с лишними скобками, что очень мешает географическому анализу



Рисунок 7 - Удаление лишних символов

Удаление символов выполнено успешно

Вывод по изученной общей информации:

* Из названий столбцов удалены пробелы
* Данные проверены на наличие дубликатов и пропусков
* Удалили лишние символы из столбца ioc\_code

# Исследовательский анализ

## 4.1 Страны, отправившие наибольшее количество участников

Отберем из начальных данных значения, которые нам пригодятся для анализа, и поместим их в отдельный DataFrame participation\_total [4].

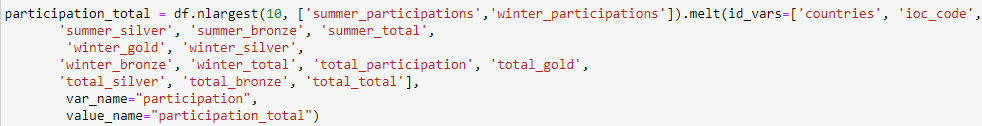


Рисунок 8 - Создание participation\_total

На основе этих данных составим график топ 10 стран, отправивших наибольшее количество участников на Олимпийские игры, разделяя зимние и летние Олимпийские игры.

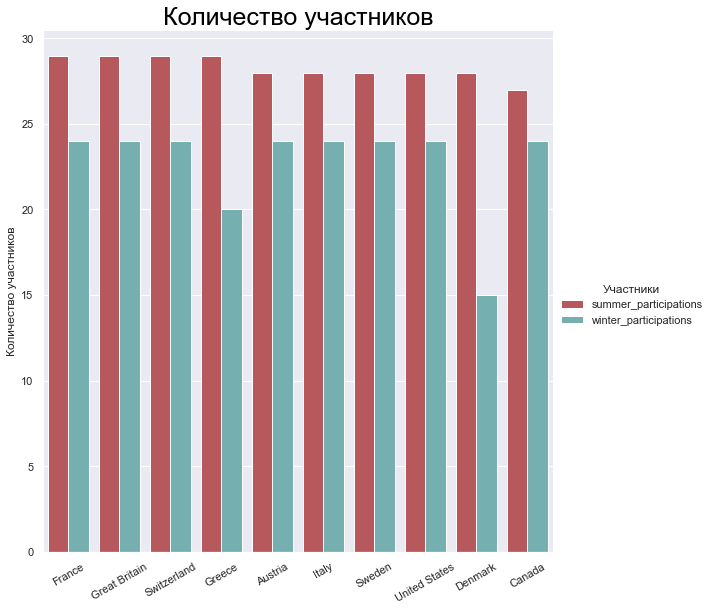


Рисунок 9 - Страны-участницы

По графику можно заметить, что на летние Олимпийские игры отправляется больше участников, это связано с большим количеством летних дисциплин. Большее количество участников приходится на страны Европы и Америки, причины этого пока что неясны.

Обратим внимание, что максимальное количество участников не превышает 30 команд.

## 4.2 Участники Олимпийских игр, нанесенные на карту

Олимпийские игры затрагивают практически весь мир. Будет интересно взглянуть как именно участники распределяются по земному шару. Для этого нанесем точками на карту страны, участвующие в Олимпийских играх, а размер точки будет зависеть от количества отправленных участников.

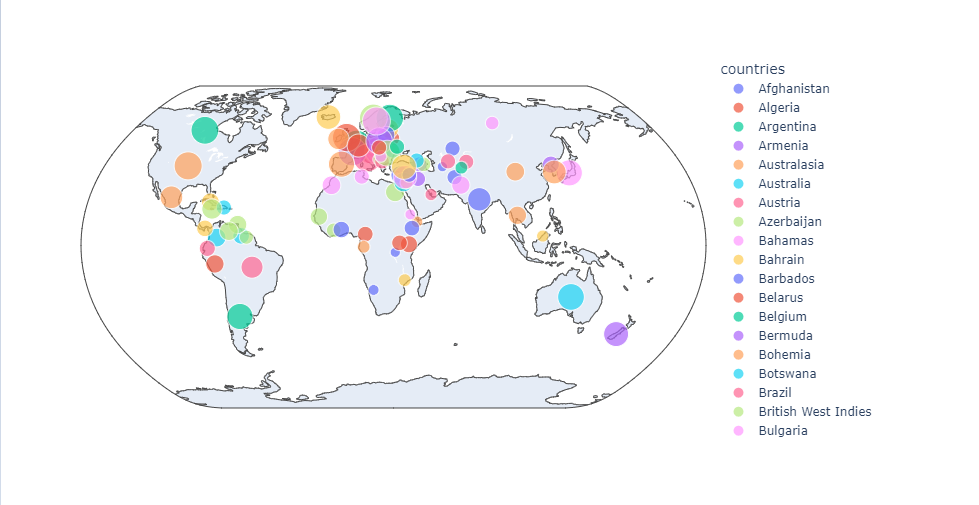


Рисунок 10 - География стран-участниц

По графику видно, что в основном страны географически расположились вокруг Европейской части мира, однако действительно много стран приезжает из Африки, Южной Америки.

## 4.3 Анализ результатов медального зачета

Для того, чтобы предсказать результаты следующих медальных зачетов, необходимо проанализировать результаты за предыдущие годы. Мы располагаем данными о результатах Олимпийских игр с самого их начала. Чтобы проанализировать результаты разделим данные на несколько таблиц: результаты летних Олимпийских игр, результаты зимних Олимпийских игр, результаты всех Олимпийских игр.

Создадим DataFrame summer\_medals из соответствующих полей, и изобразим получившийся результат на графике:



Рисунок 11 - Создание summer\_medals

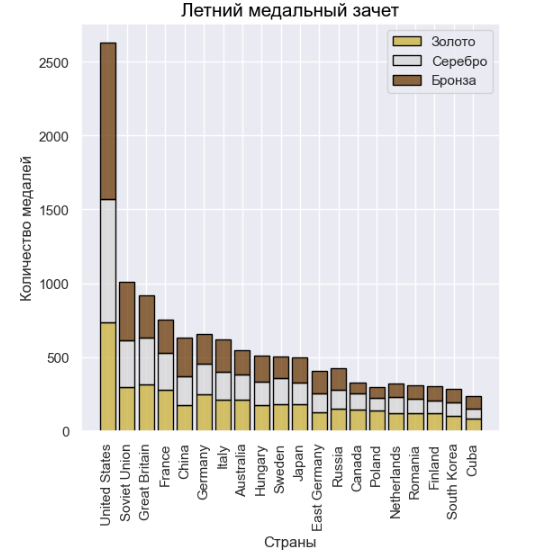


Рисунок 12 - Летний медальный зачет

Создадим DataFrame winter\_medals[2] из соответствующих полей, и изобразим получившийся результат на графике:

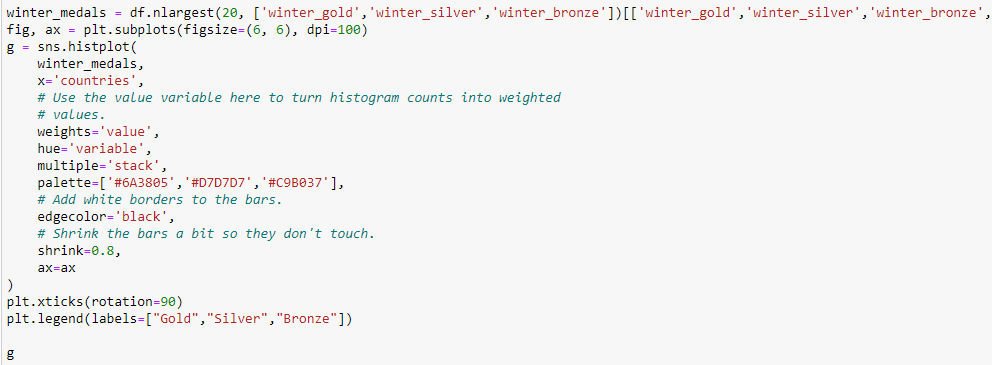


Рисунок 13 - Создание winter\_medals

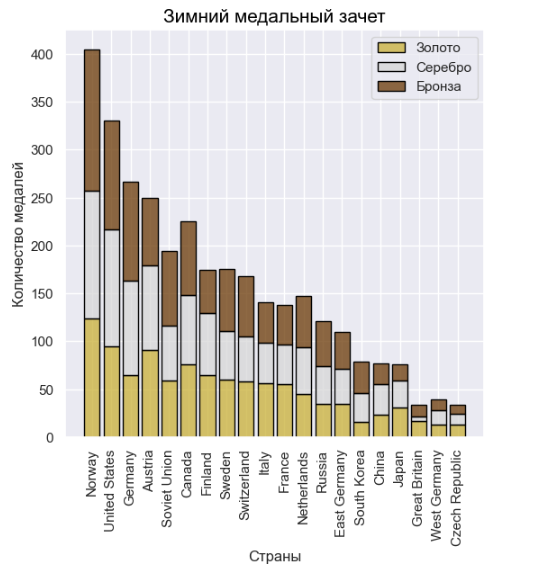


Рисунок 14 - Зимний медальный зачет

Создадим DataFrame total\_medals из соответствующих полей, и изобразим получившийся результат на графике [3]:

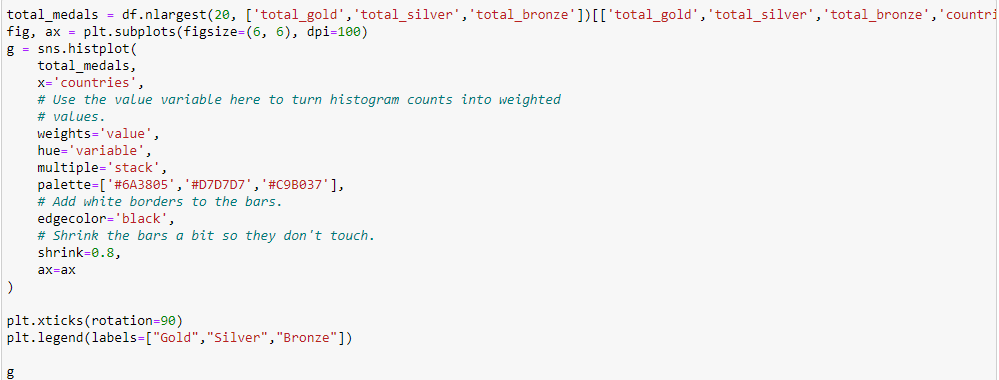


Рисунок 15 - Создание total\_medals

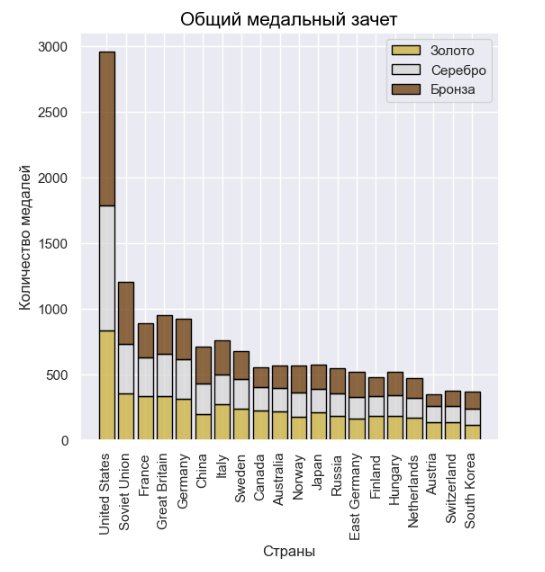


Рисунок 16 - Общий медальный зачет

Проанализируем получившиеся результаты: довольно сильно заметно, что на зимних Олимпийских играх страны северной Европы и страны северной части Америки показывают более хорошие результаты, чем на летних Олимпийских играх, и наоборот, страны южной Европы лучше показывают себя в медальных зачетах на летних Олимпийских играх.

Сильно выделяется отличное выступление США, количество медалей у которых сильно превышает значения других стран: спортсмены США отлично показывали себя и на летних, и на зимних играх. Также выделяются следующие страны: Россия, Франция, Германия, Китай. На протяжении всех Олимпийских игр они набрали наибольшее количество медалей.

К сожалению, данные не позволяют посмотреть динамику результатов каждой из стран и мы не можем предположить, находятся ли спортсмены на стадии упадка или подъема, поэтому составить точные предположения о победителях следующих Олимпийских игр невозможно. Однако наши данные и результаты предыдущих Олимпийских игр дают право нам сделать вывод, что можно ожидать эти некоторые из представленного топа стран в качестве победителей медального зачета

## 4.4 Матрица зависимости переменных

На основе предыдущих диаграмм, можно предположить, что хорошие результаты летних Олимпийских игр не гарантируют отличные результаты зимних Олимпийских игр. Такое предположение можно проверить, построив корреляционную матрицу из всех числовых значений наших данных.

Для этого создадим DataFrame numerical\_df, содержащий все числовые значения из изначальных данных[3]:



Рисунок 17 - Корреляционная матрица

Получилась довольно интересная корреляционная матрица. Наше предположение подтвердилось: результаты на летних Олимпийских играх и результаты зимних игр связаны, но корреляция не настолько большая, чтобы говорить о гарантии таких же отличных результатов. Действительно большой коэффициент корреляции получился между количеством золотых, бронзовых и серебряных медалей на летних Олимпийских играх, между количеством золотых, бронзовых и серебряных медалей на зимних Олимпийских играх, а также общее количество для летних и зимних игр золотых, бронзовых и серебряных медалей оказалось сильно связано с количеством золотых, бронзовых и серебряных медалей на летних Олимпийских играх. А вот для зимних игр такая связь оказалась слабее из-за меньшего количества дисциплин.

## 4.5 Определение и анализ эффективности выступления

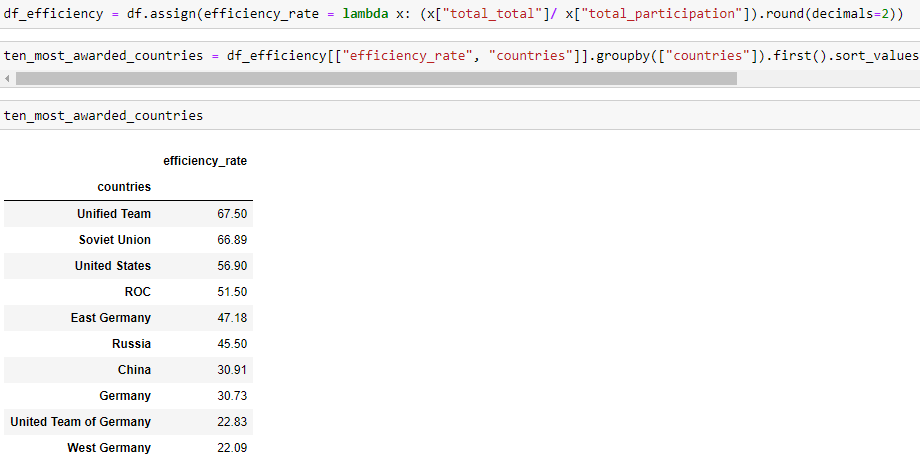
Пусть эффективность выступления — это отношение общего количества медалей на общее количество участников. Тогда чтобы проанализировать этот показатель возьмем коэффициент эффективности и объединим его с предыдущими данными : 

Рисунок 18 - Коэффициент эффективности

Найдем наиболее эффективные выступления с наибольшим рейтингом. Это имеет значение в расчете трудозатрат, и показывает высокую подготовку спортсменов. Визуализируем полученные данные с помощью диаграммы:

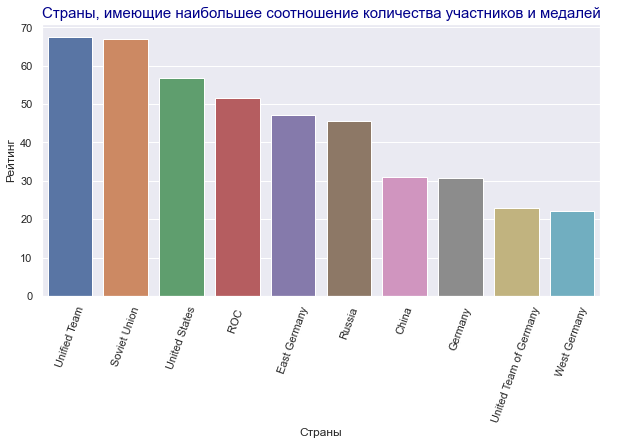


Рисунок 19 - диаграмма эффективности

# Проверка гипотез

Важным моментом принятия решения при работе с выборкой данных является использование параметрических или непараметрических статистических методов.

Параметрические статистические методы предполагают, что данные имеют известное и специфическое распределение, часто распределение Гаусса. Если выборка данных не является гауссовой, то предположения о параметрических статистических тестах нарушаются, и должны использоваться непараметрические статистические методы.

Существует ряд методов, которые можно использовать для проверки отклонения выборки данных от распределения Гаусса, называемых тестами нормальности. Поэтому первая гипотеза, которую мы будем проверять – является ли суммарное распределение медалей по странам нормальным.

Гипотеза 1:

*Нулевая гипотеза* H0: Распределение результатов медального зачета нормально

*Альтернативная гипотеза* H1*:* Распределение результатов медального зачета ненормально

Для начала попробуем проанализировать это с помощью графика распределения частот и, глядя на график, предположим, что данные не распределены нормально [1].

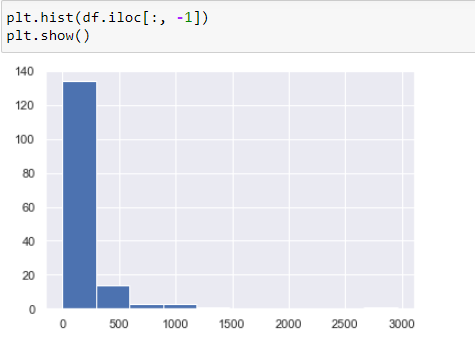


Рисунок 20 - Диаграмма частот

В подтверждение нашего заключения используем тесты нормальности: сначала тест Шапиро-Вилка, затем тест д'Агостино

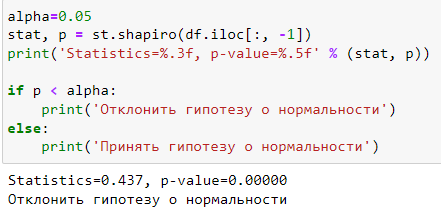
****

Рисунок 21 - Тест Шапиро-Вилка

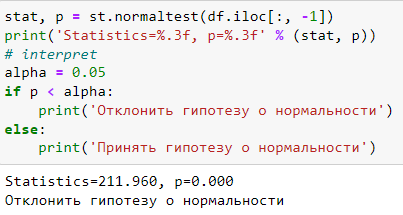
****

Рисунок 22 - Тест д'Агостино

**Вывод**

Оба теста подтвердили наше предположение, и мы можем сказать, что данные не распределены нормально.

Гипотеза 2:

**Нулевая гипотеза:** Средние результаты, показываемые Северной Европой на зимних Олимпийских играх, не отличаются от результатов остальных странах

**Альтернативна гипотеза:** Средние результаты, показываемые Северной Европой на зимних Олимпийских играх, выше, чем в остальных странах

Благодаря предыдущей гипотезе мы выяснили, что данные распределены ненормально, а значит для сравнения средних мы будем использовать непараметрический тест Манна-Уитни.

Для его применения выделим данные для стран Северной Европы: согласно классификации ООН, к странам Северной Европы относятся Ирландия, Великобритания, Норвегия, Исландия, Дания, Швеция, Финляндия.

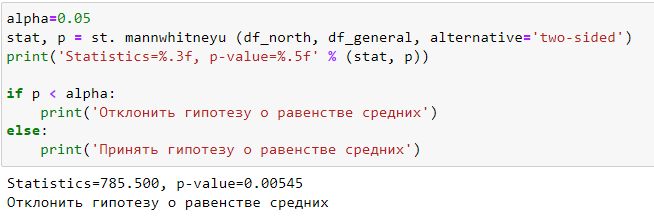


Рисунок 23 - Тест Манна-Уитни

Вывод:

Средние результаты, показываемые Северной Европой на зимних Олимпийских играх, отличаются от результатов остальных странах.

# Общий вывод

Перед анализом мы провели проверку данных на пропуски, дубликаты. Также мы проверили названия столбцов и обнаружили в них ошибки. Удалили лишние символы из столбца с кодом страны.

В ходе анализа мы выяснили:

* + 1. Какие страны отправили больше всех участников;
    2. Кто победил в медальном общем медальном зачете, летнем и зимнем зачете;
    3. Расположение стран-участниц на географической карте;
    4. Какие страны выступали наиболее эффективно;
    5. Взаимосвязь между переменными.

Также мы сформулировали и проверили гипотезы, получилось, что:

1. Данные о наградах не распределены нормально
2. Средние результаты, показываемые Северной Европой на зимних Олимпийских играх, выше, чем в остальных странах

# ЗАКЛЮЧЕНИЕ

В ходе научно-исследовательской работы был проведен анализ базы данных «Анализ результатов олимпийских игр с 1896 года», выявлены важные статистические данные, которые могут помочь в лучшем понимании следующих результатов Олимпийских игр.

Для выполнения данной работы было изучено новое программное обеспечение: «Pandas» и «Seaborn» на базе языка программирования Python.

В ходе анализа мы выяснили:

* + 1. Какие страны отправили больше всех участников;
    2. Кто победил в медальном общем медальном зачете, летнем и зимнем зачете;
    3. Расположение стран-участниц на географической карте;
    4. Какие страны выступали наиболее эффективно;
    5. Взаимосвязь между переменными.

# СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Лекции «Оперативный анализ данных», преподаватель кафедры ИУ5 Маслеников К.Ю.
2. Методические указания по программному обеспечению «Pandas»
3. Методические указания по программному обеспечению «Seaborn»
4. «Pandas. Работа с данными» (2020), Автор: Абдрахманов М. И.
5. «Python. Визуализация данных: Matplotlib, Seaborn, Mayavi»