**МИНОБРНАУКИ РОССИИ**

**Федеральное государственное автономное образовательное  
 учреждение высшего образования   
«Южный федеральный университет»**

**Институт высоких технологий и пьезотехники**

****

**Кафедра прикладной информатики и инноватики**

**Направление: 09.03.03 "Прикладная информатика"**

**Большие данные  
Отчёт по проекту**

**«Анализ футбольного рынка»**

Выполнили студенты 3 курса 7 группы

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Колисниченко М.В.

подпись

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Сылкин А.А.

подпись

Проверил старший преподаватель

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Яценко Д.В.

подпись

**Ростов-на-Дону, 20****24**

СОДЕРЖАНИЕ

[**1.** **Описание проблемы** 3](#_Toc169543549)

[**2.** **Постановка задачи** 3](#_Toc169543550)

[**3.** **Описание набора данных** 4](#_Toc169543551)

[**4.** **Ход работы** 4](#_Toc169543552)

[**4.1. Предобработка данных** 4](#_Toc169543553)

[**4.2. Получение значений** 10](#_Toc169543554)

[**4.3. Анализ и визуализация результатов** 11](#_Toc169543555)

[**5.** **Выводы** 14](#_Toc169543556)

# **Описание проблемы**

Существует тенденция, что в разных национальных чемпионатах для того, чтобы у игрока поднялась стоимость на m миллионов евро, ему нужно сделать разное количество полезных действий. Другими словами, 1 полезное действие в зависимости от лиги, где играет футболист, имеет разную стоимость.

Например, Английская Премьер-Лига (АПЛ) считается одной из богатых лиг, и бытует мнение, что ценники на футболистов в чемпионате совсем не соизмеримые с тем, что делает игрок на поле.

# **Постановка задачи**

При помощи полученных во время учёбы знаний, а также дополнительного материала, изученного при использовании различной литературы, создать решение, по которому можно проанализировать, как меняется стоимость футболистов от различных факторов, и решить тем самым следующие подзадачи:

* выяснить какие данные нужны и полезны;
* преобразовать данные к необходимому виду;
* обучить линейную регрессию, дав на вход преобразованные данные;
* получить коэффициенты регрессии нужных признаков;
* визуализировать результаты.

Поскольку выбранный набор данных имеет полезные действия атакующего характера – голы и голевые передачи, было решено рассмотреть случаи роста цен у нападающих и полузащитников.

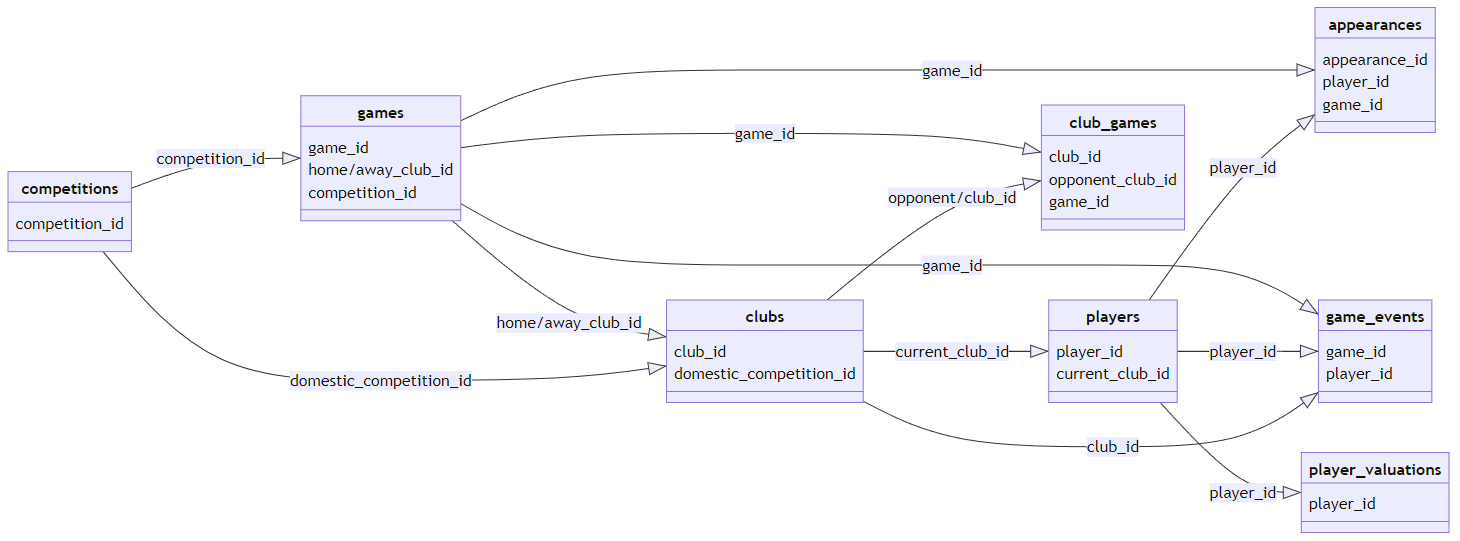
Для объективности данных было сделано решение рассмотреть ценообразования игроков с сезона 2017/2018. В этом сезоне нападающий «Барселоны» Неймар перешёл во французский «ПСЖ» за рекордные в наши дни 222 млн. евро, после чего цены на рынке существенно изменились.

Рассмотренные футбольные лиги стали – классический топ-5 (Англия, Испания, Италия, Германия, Франция) и Россия с Турцией.

Было решено использовать учебный кластер и PySpark для обработки данных.

# **Описание набора данных**

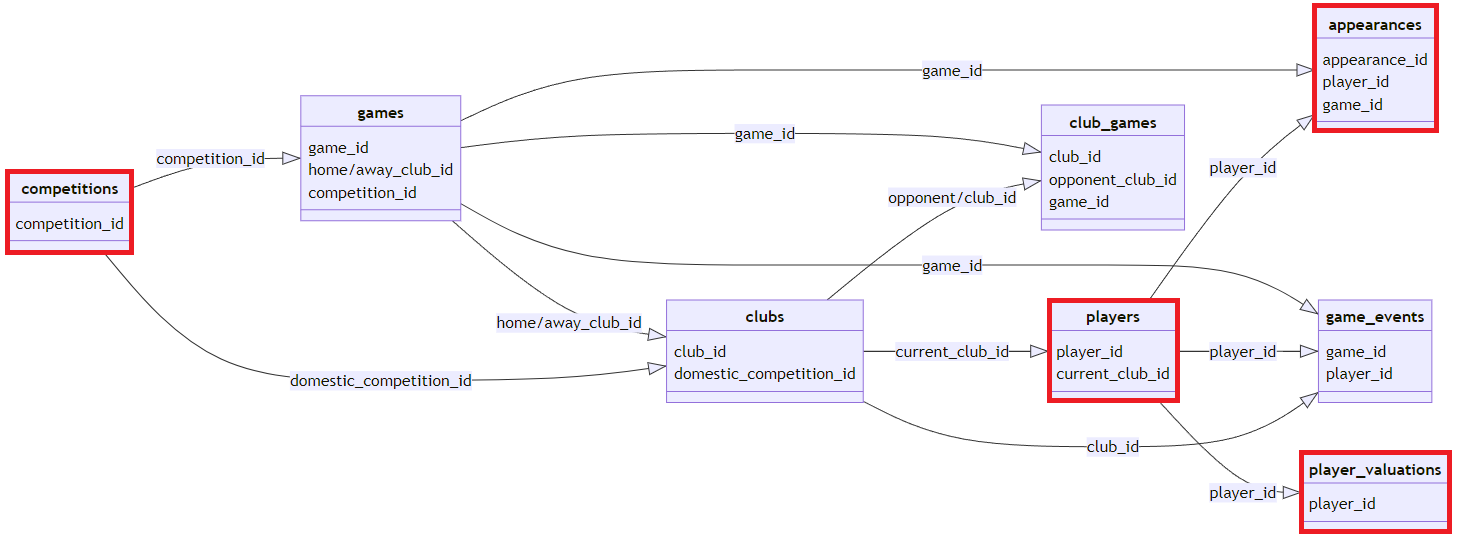
Набор данных **Football Data from Transfermarket** состоит из нескольких CSV-файлов с информацией о соревнованиях, играх, клубах, игроках и выступлениях, которая автоматически обновляется раз в неделю. Каждый файл содержит атрибуты сущности и идентификаторы, которые можно использовать для их объединения.



# **Ход работы**

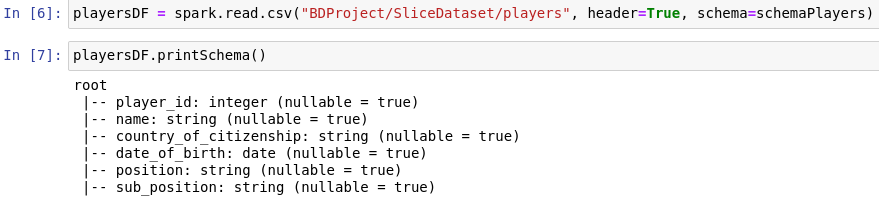
# **4.1. Предобработка данных**

1. Исходя из поставленной задачи было выяснено, что наиболее нужная и полезная информация хранится в CSV-файлах об игроках, их стоимости в разные периоды времени, их появления на поле и соревнования, в которых они участвуют.

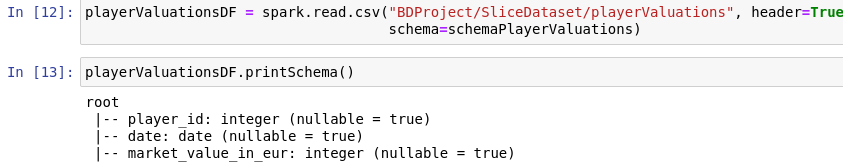


При помощи Spark Dataframe API выгрузили соответствующие файлы и назначили схему с интересующими нас признаками (столбцами).

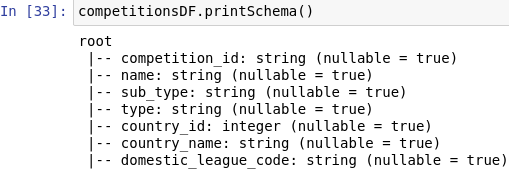
1) players:



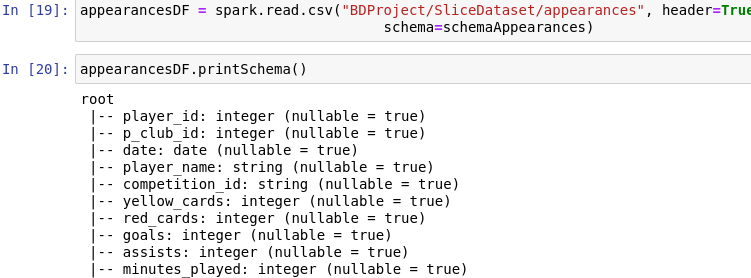
2) player\_valuations:



3) competitions:

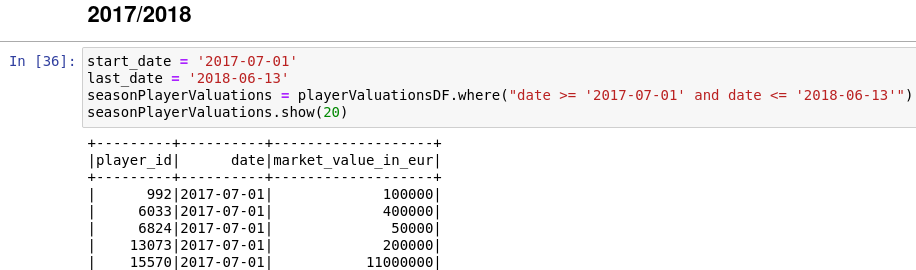


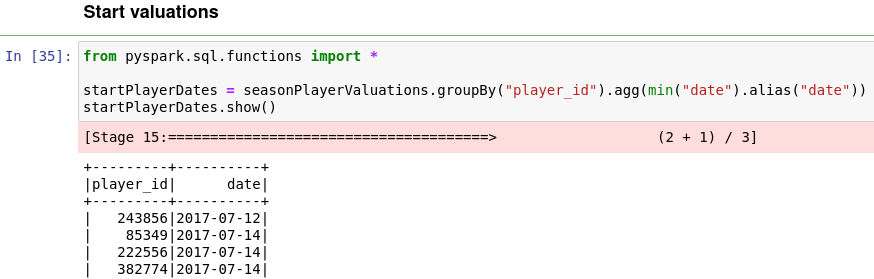
4) appearances:

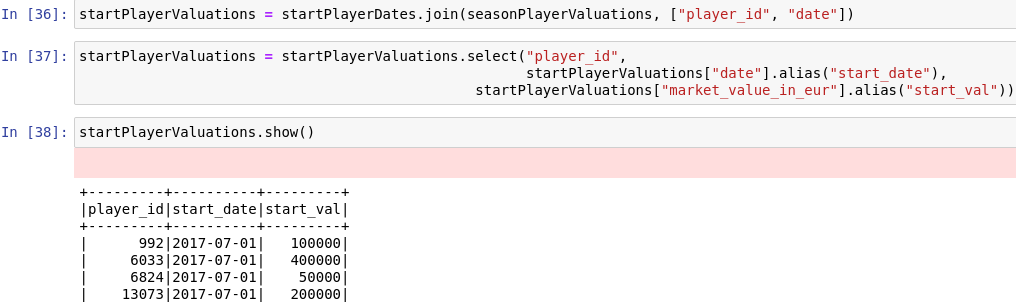


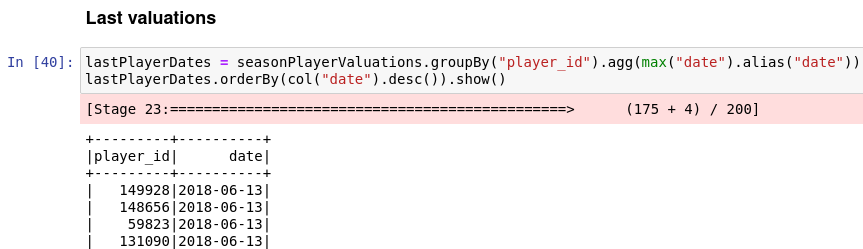
**2. Работа с playerValuationsDF**

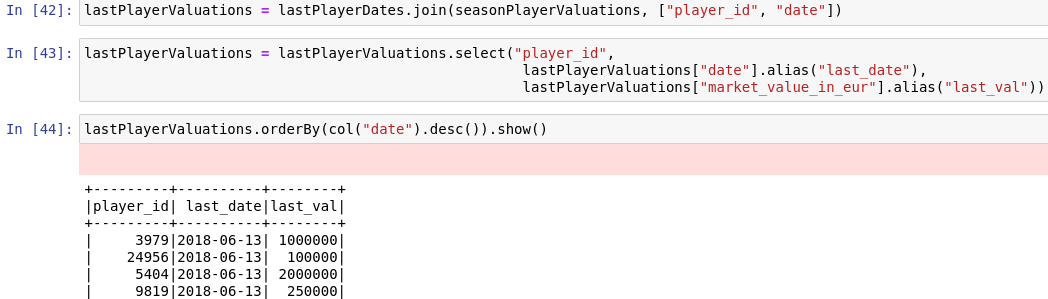
Преобразовали playerValuationsDF таким образом, чтобы он отображал информацию о начальной стоимости и конечной стоимости в конкретном сезоне.



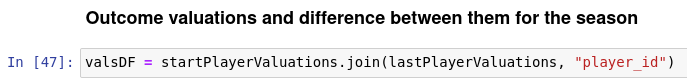


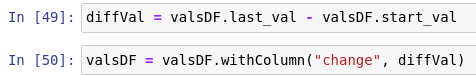


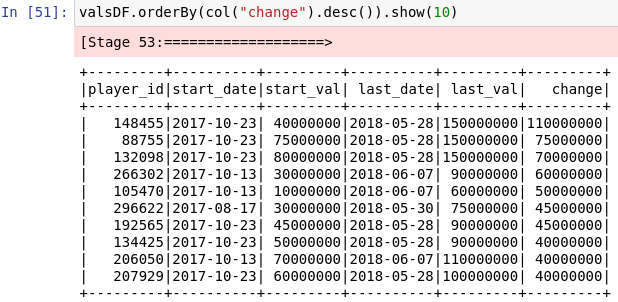




Вычислили разницу между конечной и стартовой стоимостью в сезоне тем самым, получив рост (положительный или отрицательный) в цене за игровой год.



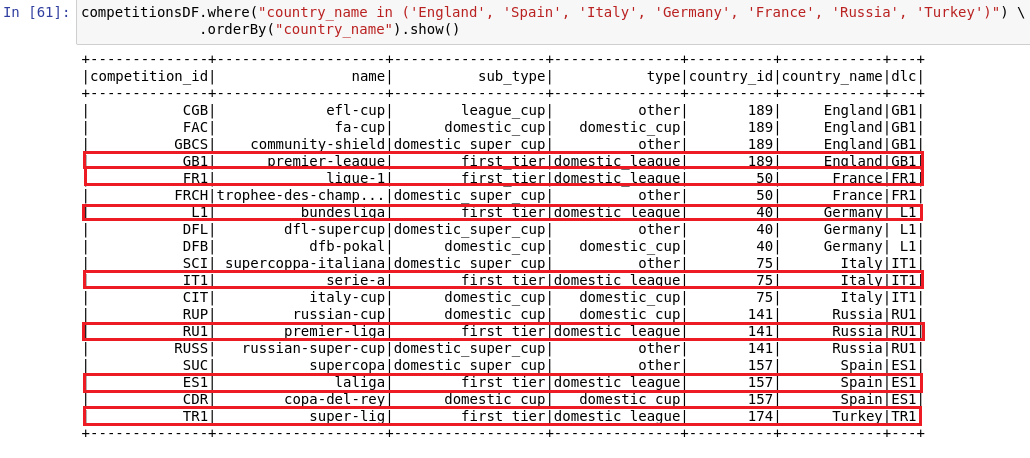




**3. Работа с competitionsDF**

Выяснили, какие competition\_id имеют интересующие нас лиги.

England – «GB1», Spain – «ES1», Italy – «IT1», Germany – «L1», France – «FR1», Russia – «RU1», Turkey – «TR1».

****

**4. Работа с appearancesDF**

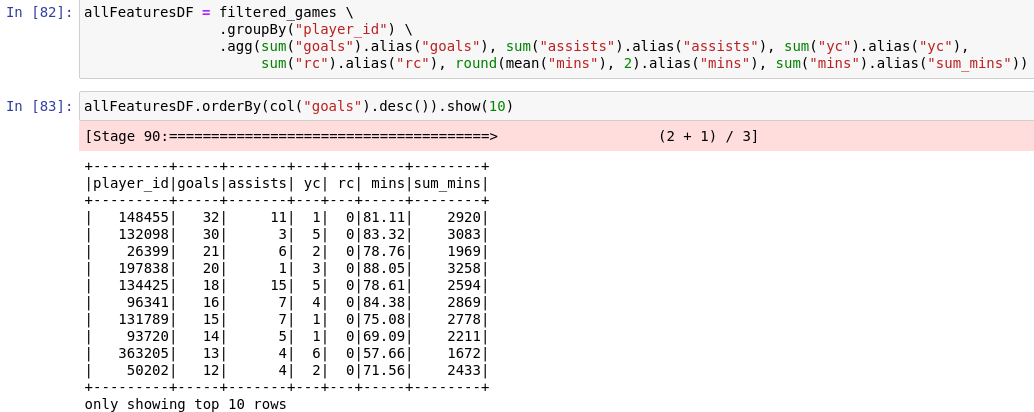
Преобразовали appearancesDF в соответствии с сезоном и лигой.

cid – competition\_id, который был найден выше, date – дата, которая фиксирует результативное действие игрока.

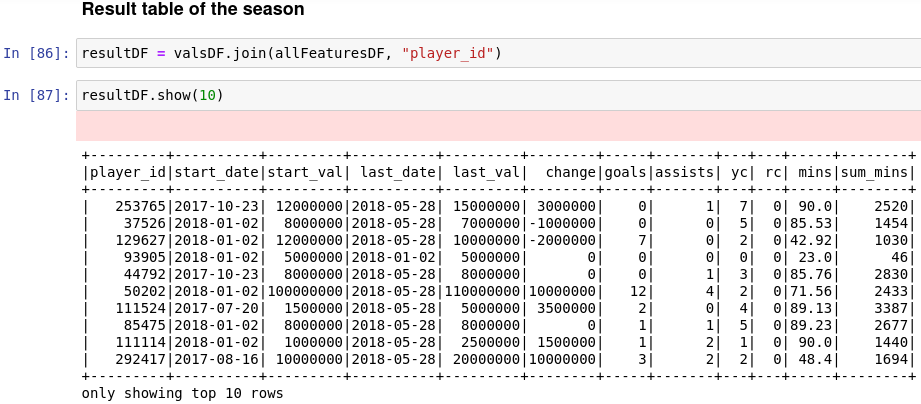


Сделали группировку по player\_id. Таким образом выяснили, сколько результативных действий делает игрок за сезон.

Goals – голы, assists – голевые передачи, yc – жёлтые карточки, rc – красные карточки, mins – среднее количество минут, sum\_mins – суммарное количество минут.

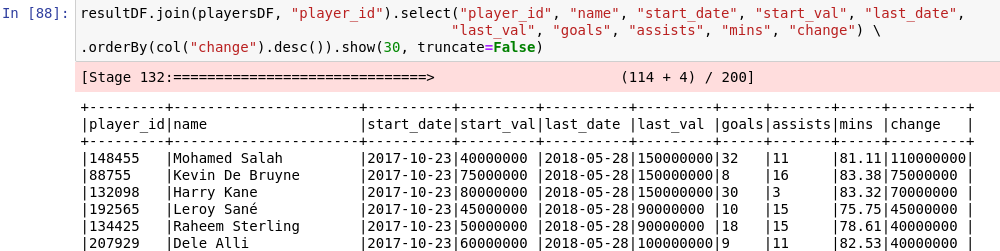


**5. Объединение двух результирующих DF в один**



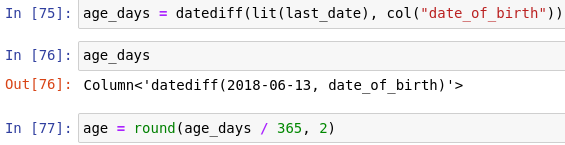
**6. Проверка правильности результатов**

Объединили результирующий DF с playersDF для проверки реальности данных.



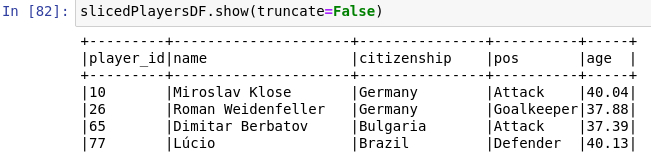
**7. Работа с playersDF**

Вычислили разницу между датой рождения игрока и датой сезона тем самым, получив возраст игрока на конкретный сезон.



Использовали dropna, т.к. у некоторых игроков отсутствует дата рождения. Выбрали нужные столбцы: country\_of\_citizenship – паспорт, position – общая позиция игрока, age – возраст.





8. Было сделано предположение, что на стоимость также влияют полезные действия в еврокубках, потому **пункты 3 и 4** были повторены для Лиги Чемпионов и Лиги Европы (1-й и 2-й по значимости клубные международные турниры).

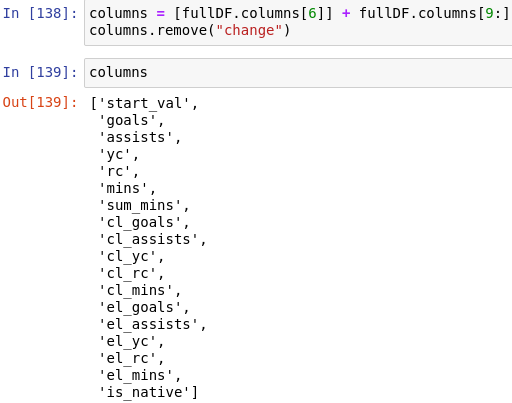
9. Был написан общий скрипт с циклами описанных выше действий для выбранных лиг, сезонов, возрастов и позиций.



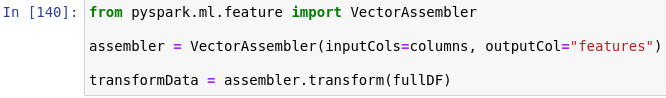
В итоге у нас вышло 42 DF’а (7 стран \* 3 возрастные группы \* 2 позиции).

# **4.2. Получение значений**

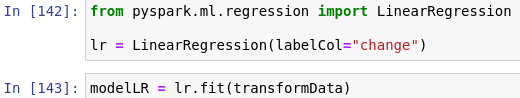
1. Создали список, содержащий интересующие названия признаков. Исключили «change», т.к. является меткой.



2. Создали преобразователь, чтобы собрать столбцы объектов в виде Vector  
Column. Чтобы выполнить это воспользовались VectorAssembler().

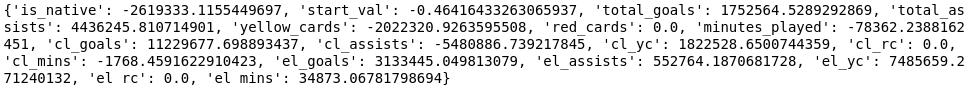


3. Использовали класс LinearRegression для создания оценщика линейной регрессии.



4. Получили значения **коэффициентов** линейной регрессии.





Они определяют средние изменения результата с изменением соответствующего фактора на 1 при неизменном значении других факторов, зафиксированных на среднем уровне.

5. Оценили качество построенной модели с помощью коэффициента детерминации.





**Коэффициент детерминации** оценивает, какая доля вариации зависимой  
переменной объясняется независимыми переменными модели. Значение R²  
близкое к 1 указывает на хорошее качество модели

# **4.3. Анализ и визуализация результатов**

1. Написали общий скрипт описанных действий в **4.2.** для полученных ранее DF’в.

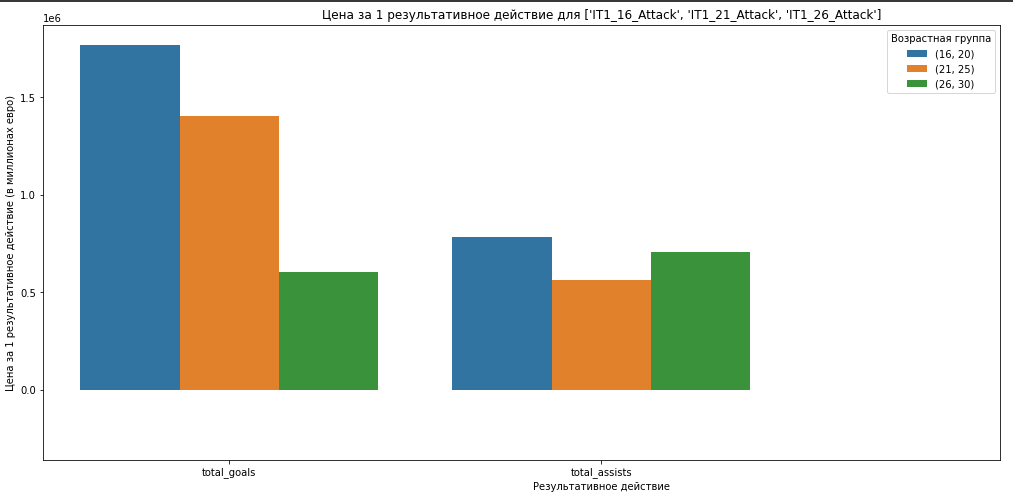
2. При помощи библиотеки питона matplotlib построили столбчатые диаграммы 3-х видов:

1) стоимость голов и голевых передач по позиции с разными возрастами в рамках одной лиги;

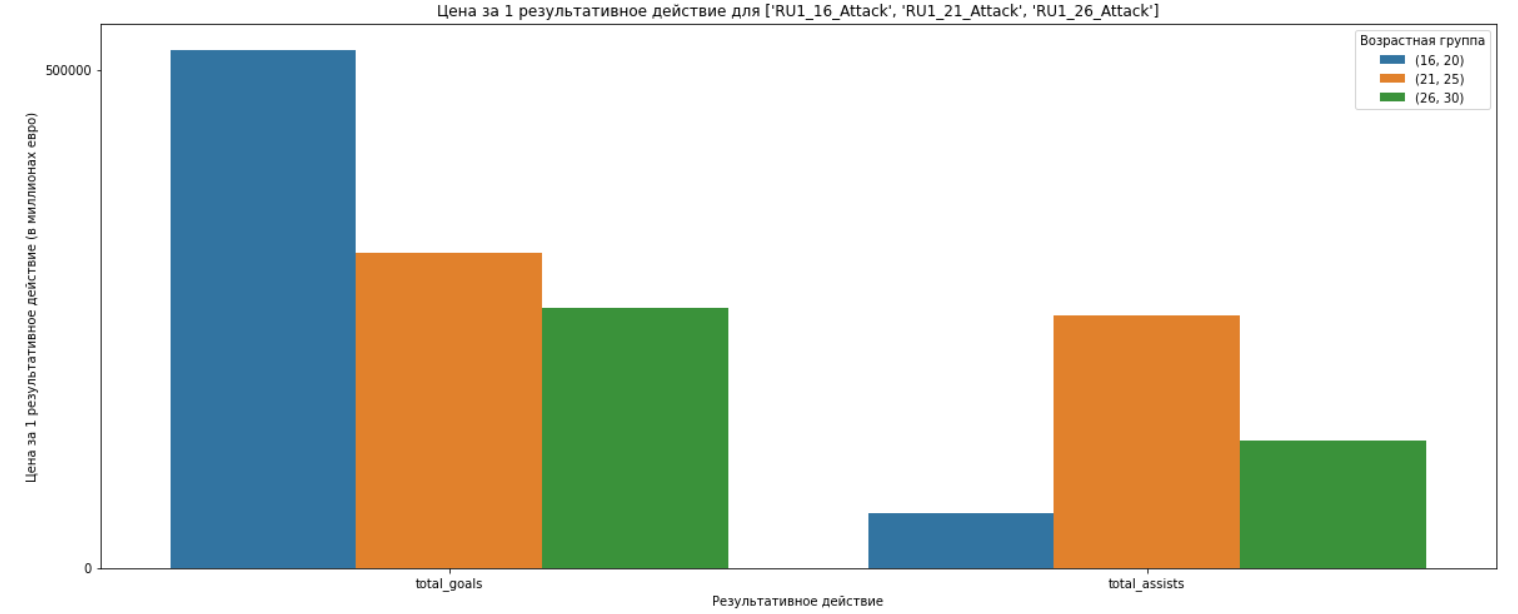
2) стоимость голов и голевых передач игроков разных позиций в рамках одной лиги;

3) стоимость голов и голевых передач игроков в разных лигах.

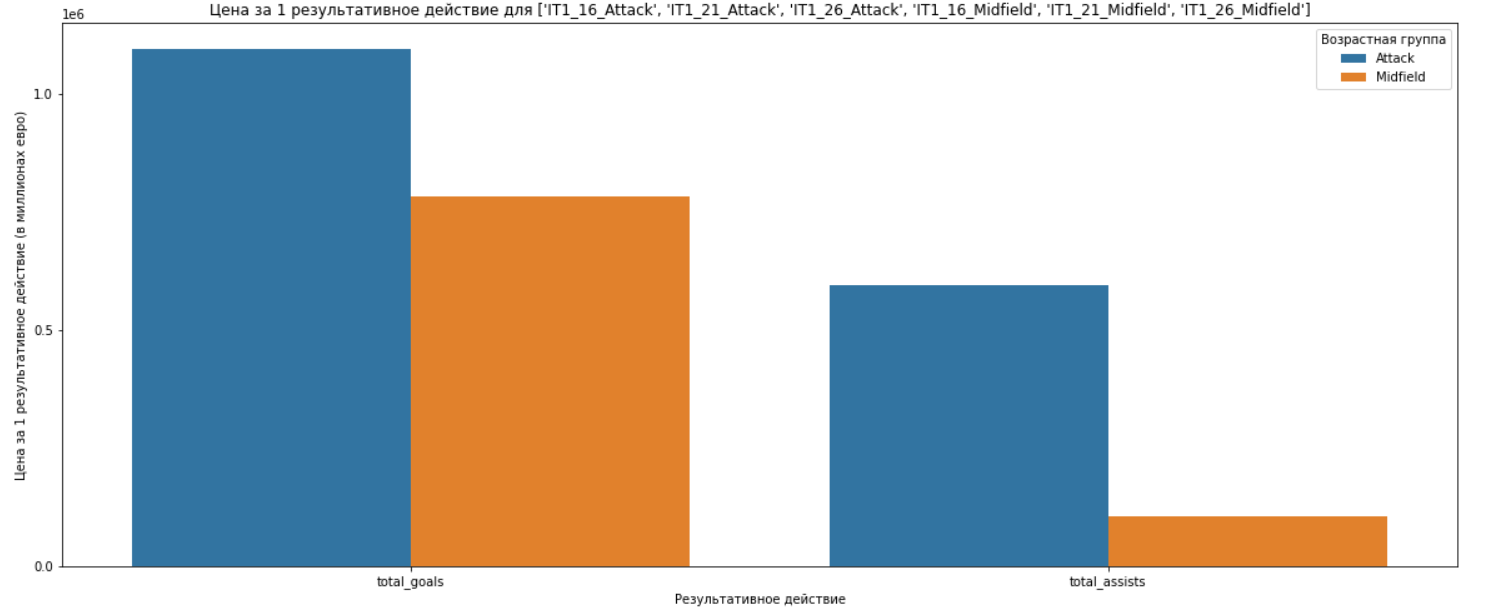
**Пример первой диаграммы в итальянской лиге для нападающих:**



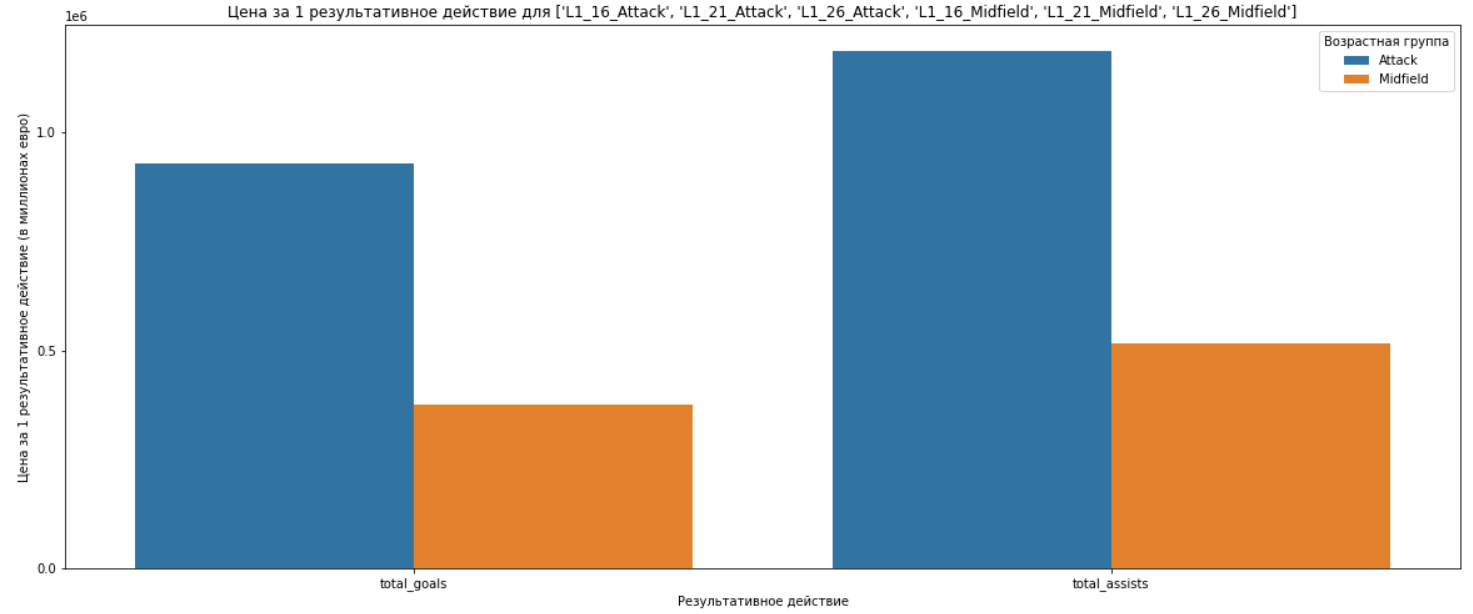
**Пример первой диаграммы в российской лиге для нападающих:**



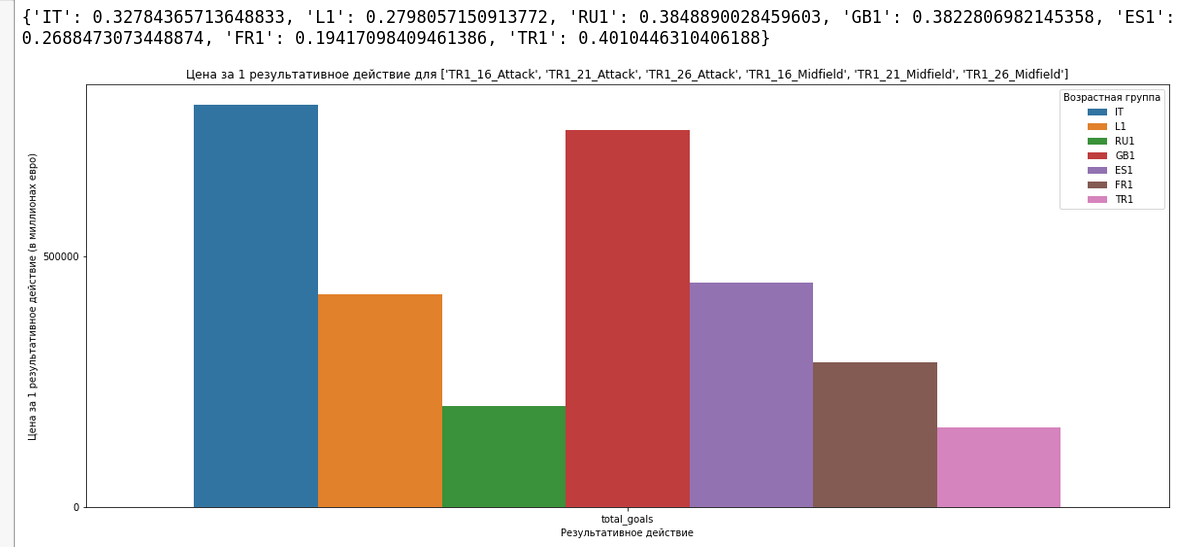
**Пример второй диаграммы в итальянской лиге для нападающих и полузащитников:**

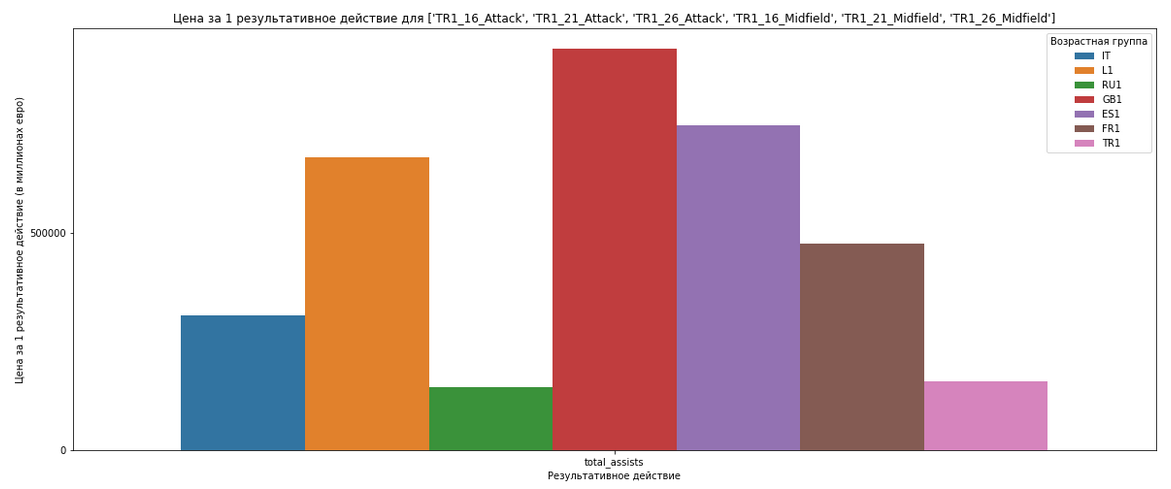


**Пример второй диаграммы в немецкой лиге для нападающих и полузащитников:**



**Третий тип диаграммы:**

****

****

# **Выводы**

Исходя из результатов и соответствующим им графикам, можно убедиться в том, что главный тезис футбольных болельщиков частично верный – в АПЛ результативное действие стоит дороже, чем в других национальных чемпионатах. Однако, можно увидеть и то, что в итальянской Серии А также дорого ценятся голы.

Также можно сделать вывод о том, что возраст и позиция также влияет на рост цены: чем моложе футболист – тем дороже его гол/передача, и аналогично с позициями нападающего и полузащитника.