|  |  |
| --- | --- |
| Gerb-BMSTU_01 | **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 и системы управления»**

**Кафедра ИУ5 «Системы обработки информации и управления»**

Отчет по рубежному контролю №1

Разведочный анализ и подготовка данных

по дисциплине «Машинное обучение»

Вариант №2

Выполнил:

Кондратьев Максим

19.04.2019

Проверил:

к.т.н., доц., Ю. Е. Гапанюк

подпись, дата

Москва, 2019 г.

Задача:

Для заданного набора данных проведите обработку пропусков в данных для одного категориального и одного количественного признака. Какие способы обработки пропусков в данных для категориальных и количественных признаков Вы использовали? Какие признаки Вы будете использовать для дальнейшего построения моделей машинного обучения и почему?

Датасет:

https://www.kaggle.com/drgilermo/nba-players-stats - статистика игроков за 67 сезонов NBA.

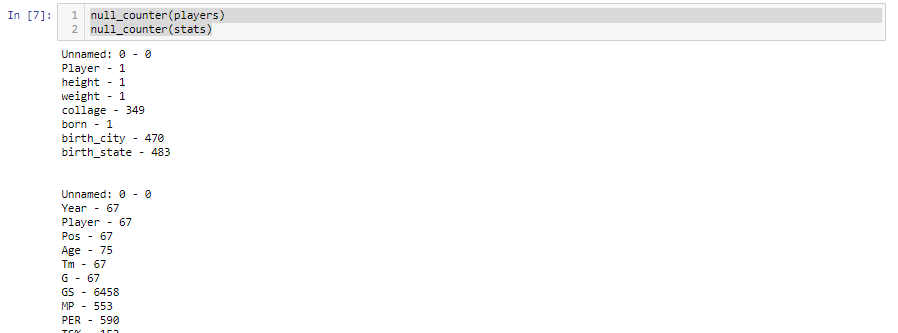
Импортирую необходимые модули, загружаю данные из датасета, хранящиеся в формате csv в Pandas DataFrame:



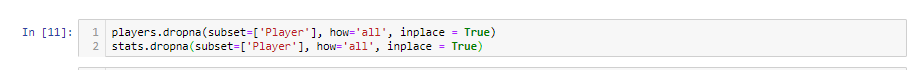
Датасет состоит из двух основных файлов. Оба содержат категориальные и числовые фичи:



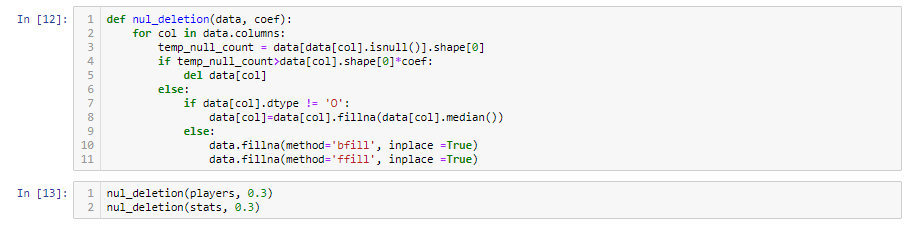
Оба файла имеют пропуски в данных и в категориальных, и в числовых полях:



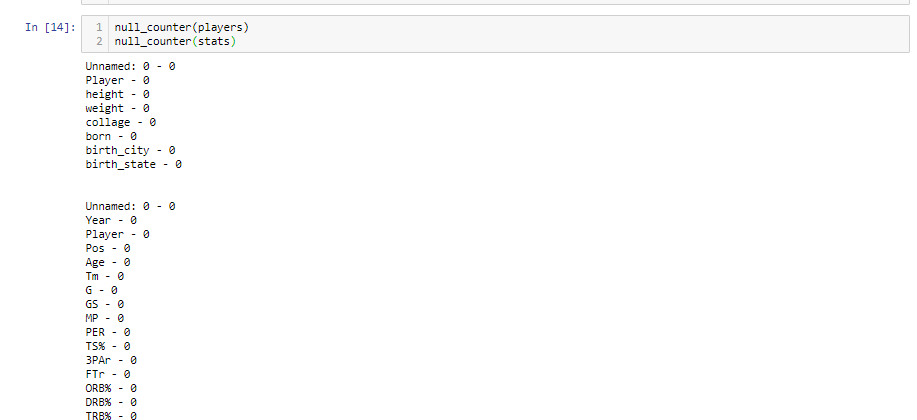
Перед заполнением пропусков, удалю строки, не содержащие значения в поле имени игрока, так как данная фича является основой для объединения двух файлов (внешний ключ).



Для заполнения пропусков использую написанную мной функцию «nul\_deletion». Согласно алгоритму данной функции, все столбцы проверяются на процент пропусков. Если пропусков в данных больше задаваемого коэффициента, столбец полностью удаляется. В случае меньшего число пропущенных данных, в зависимости от типа столбца, пропуск заполняется либо медианой существующих значений для числовых фич, либо значением под ним, а если такого нет - над ним.



Все пропуски успешно заполнены:



Для дальнейшего построения модели я буду использовать признаки, выбранные на основе корреляции. Признаки, необходимые для построения модели должны слабо коррелировать между собой, при этом иметь большую корреляцию с целевым признаком. На основе этого правила я выбрал ряд фич представленных на корреляционной матрице с их описанием:

