нетология

# Дипломная работа

## Анализ данных на основе результатов матчей Английской Премьер Лиги с 2014 по 2021 гг

Курс: Аналитик данных DAU-28

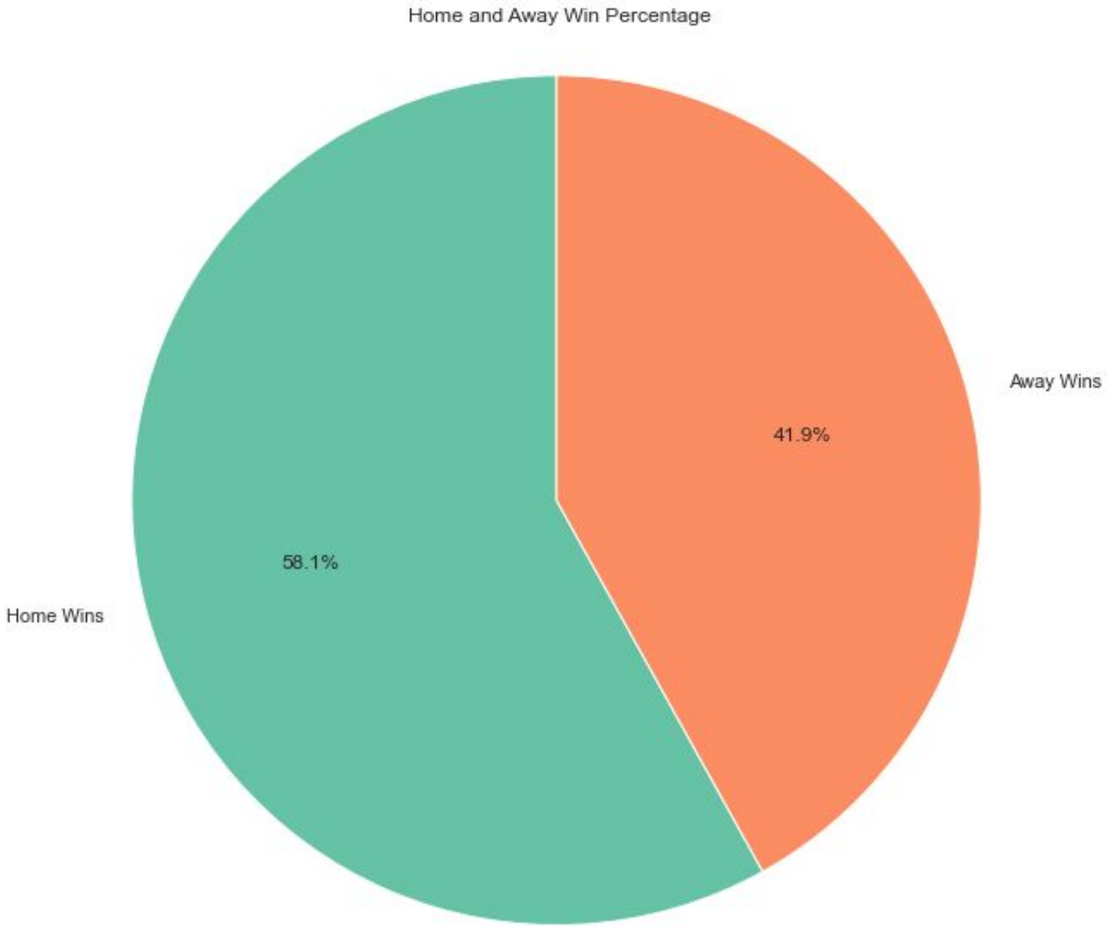
# Постановка задачи

Цель данной работы заключается в проведении анализа данных матчей Английской Премьер Лиги за период с 2014 по 2021 гг., а также проверке гипотез, основанных на этих данных.

# Основные гипотезы для проверки

- Существует сильная корреляция между ожидаемыми голами (xG) и фактическим количеством голов, забитых командами в матчах.
- Команды, которые зарабатывают больше угловых ударов в матче, чаще выигрывают или набирают больше очков.
- Команды, которые ведут после первого тайма, чаще выигрывают матчи или набирают больше очков.
- Команды с более высоким количеством желтых и красных карточек чаще теряют очки в матчах из-за снижения эффективности игры.
- Команды, играющие на домашнем стадионе, побеждают чаще, чем в гостях.
- Команды, которые наносят большее количество ударов в створ ворот чаще выигрывают.

# Команды играющие на домашнем стадионе побеждают чаще, чем в гостях.

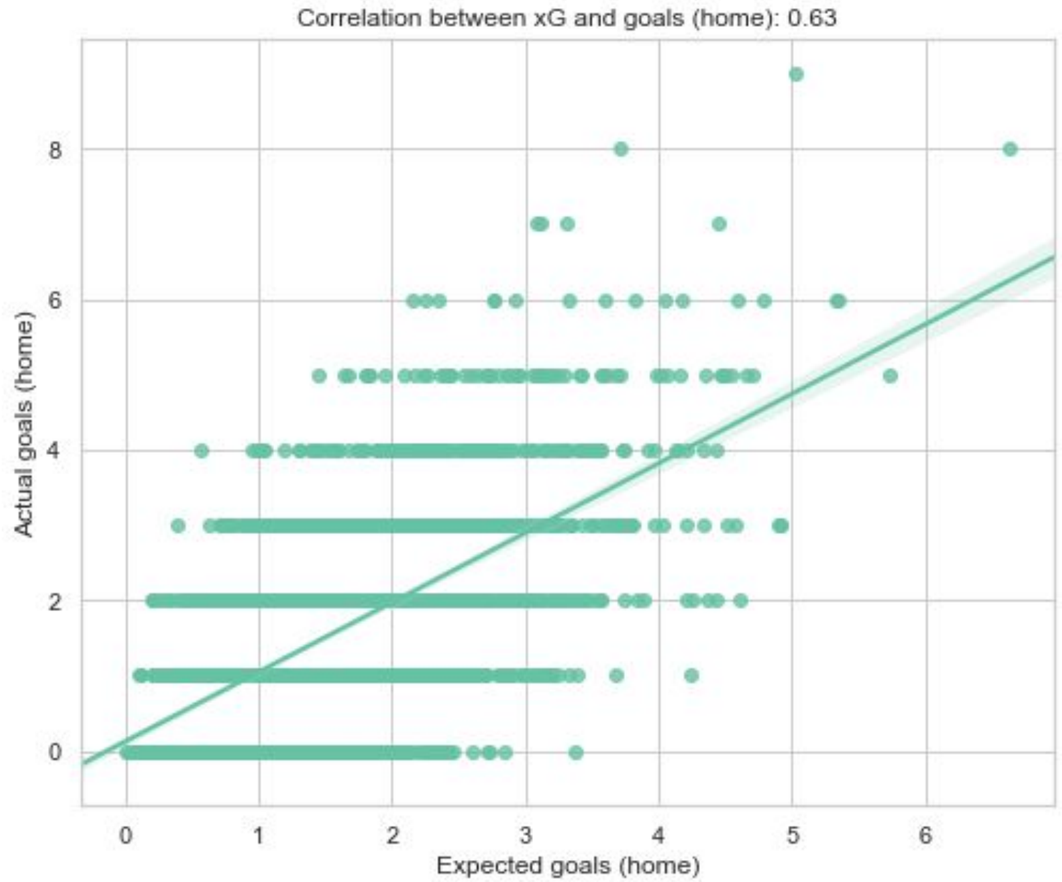


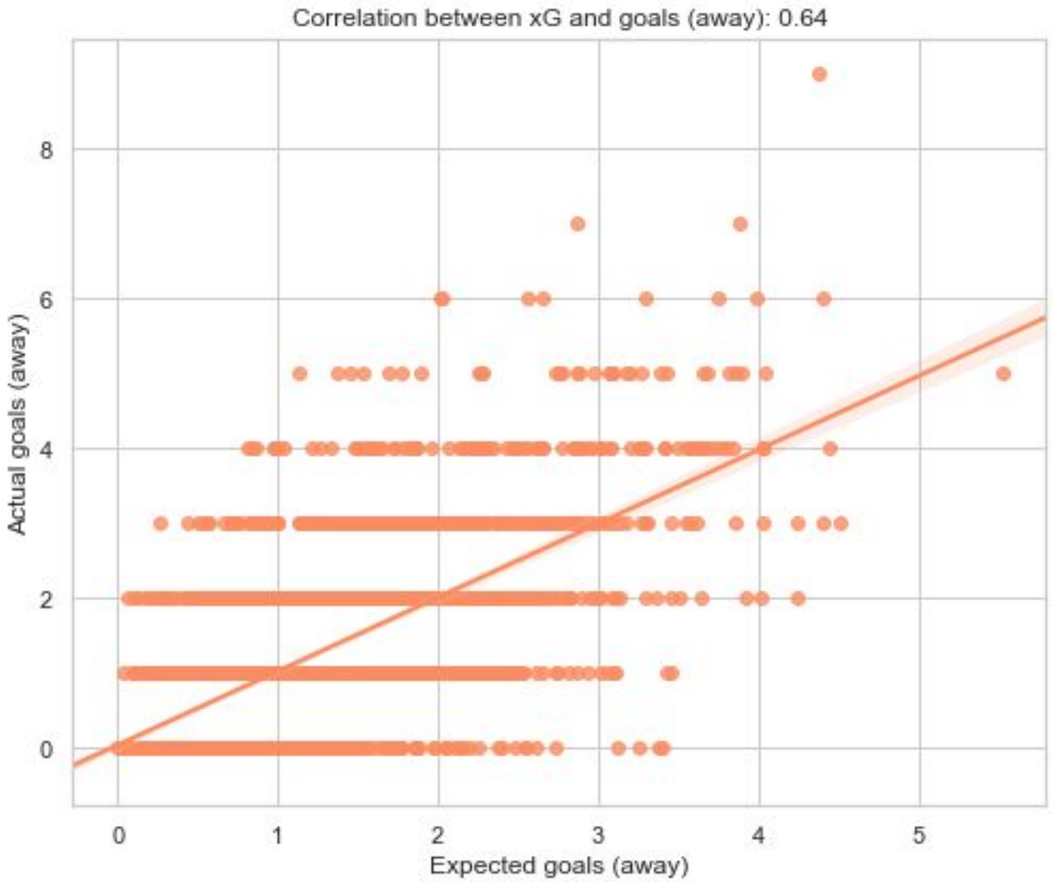
# Выводы:

- В большинстве сезонов с 2014 по 2021 год домашние команды в АПЛ чаще всего одерживали победы, что свидетельствует о том, что домашний стадион и поддержка болельщиков могут оказывать положительное влияние на результаты команд.

- В 2020 году произошло отклонение от общей тенденции - такое изменение в результатах может быть связано с влиянием пандемии COVID-19 на спортивные мероприятия и состояние команд. Возможно, отсутствие или ограничение числа зрителей на стадионах снизило преимущество домашних команд, что привело к росту проигрышей

# Существует сильная корреляция между ожидаемыми голами (xG) и фактическим количеством голов, забитых командами в матчах.



# Выводы:

- Коэффициенты корреляции 0.63 и 0.64 указывают на среднюю положительную корреляцию между ожидаемыми голами (xG) и фактическим количеством голов для домашних и гостевых команд. Это означает, что в целом команды, имеющие более высокий показатель xG, часто забивают больше голов.

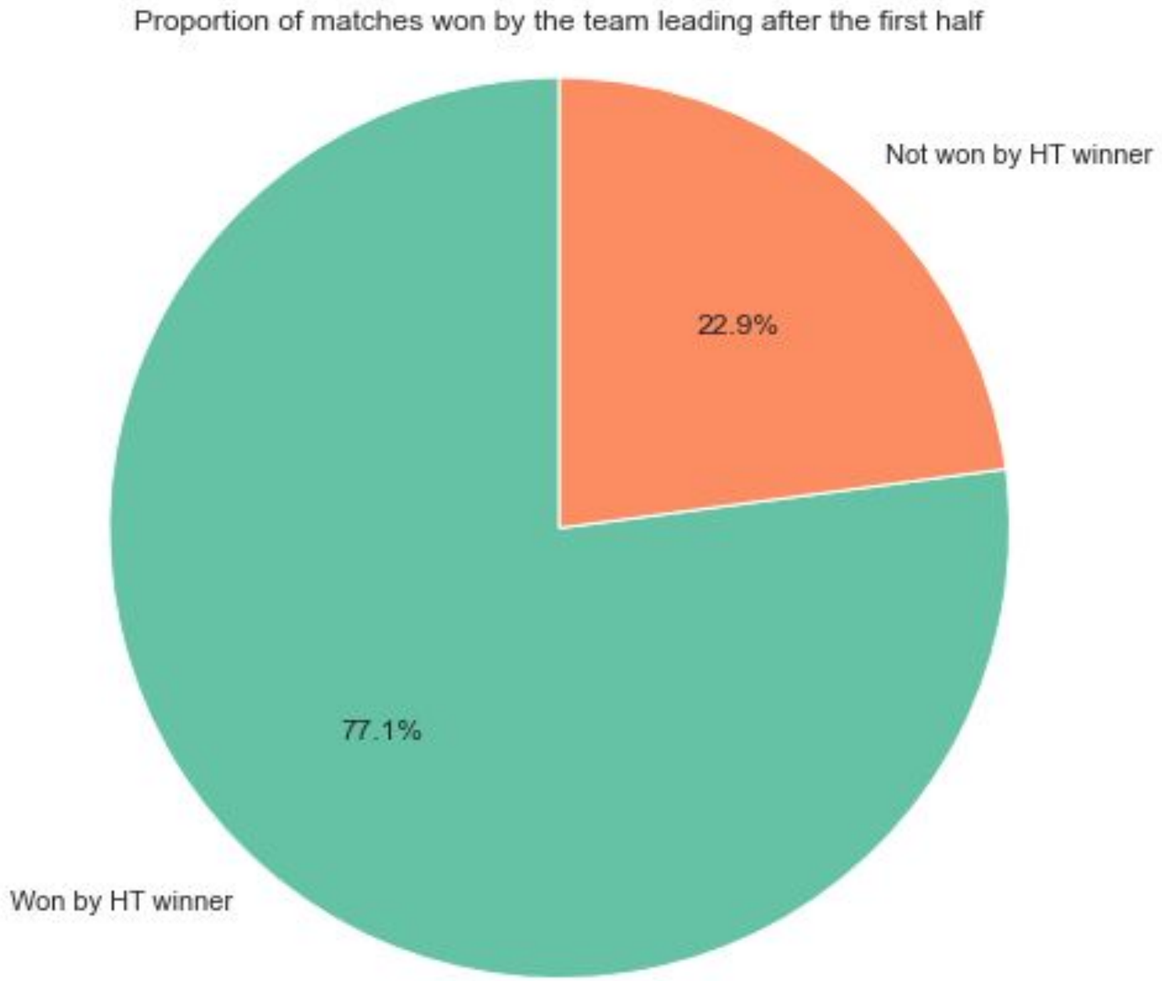
- Полученные результаты подтверждают гипотезу о существовании корреляции между ожидаемыми голами (xG) и фактическим количеством голов, забитых командами в матчах. Однако стоит отметить, что корреляция средняя, а не сильная.

# Команды, которые зарабатывают больше угловых ударов в матче, чаще выигрывают или набирают больше очков.

**Выводы:**

- Доля побед команды с большим количеством угловых ударов составляет примерно 38%.
- Коэффициент корреляции  между числом угловых ударов и количеством забитых голов для домашних матчей составляет 0.018.
- Коэффициент корреляции между числом угловых ударов и количеством забитых голов для  выездных матчей составляет -0.0047.
- На основе этих результатов можно сделать вывод, что количество угловых ударов не имеет существенного влияния на результаты матчей, как для команд, играющих дома, так и для команд, играющих на выезде.

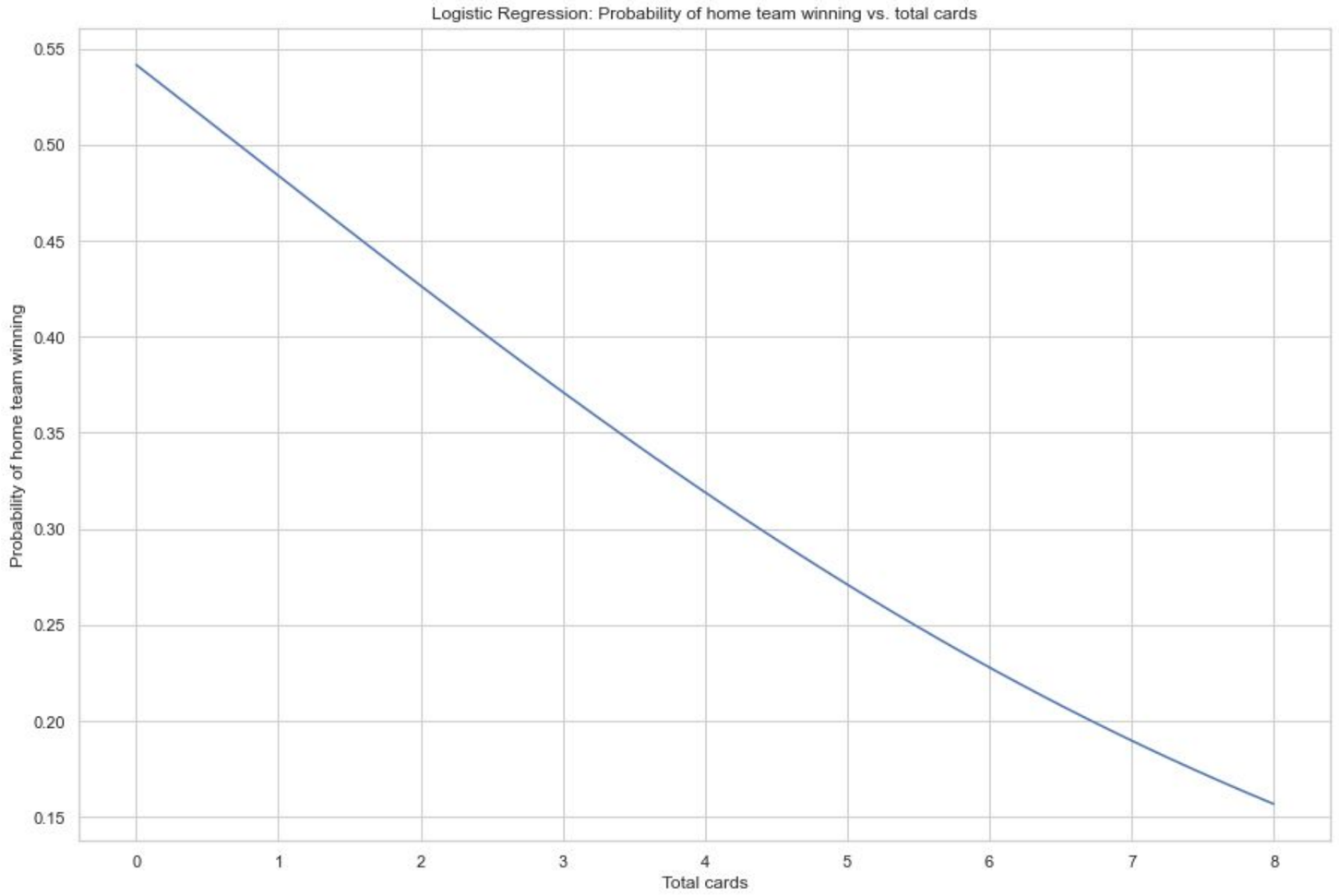
# Команды, которые ведут в счете после первого тайма, чаще выигрывают матчи или набирают больше очков.



# Выводы:

- Доля матчей, выигранных командой, которая ведёт в счете после первого тайма составляет примерно 77.1%. Исходя из этого, можно сделать вывод, что команды, которые ведут после первого тайма, действительно чаще выигрывают матчи.

# Команды с более высоким количеством желтых и красных карточек чаще теряют очки в матчах из-за снижения эффективности игры



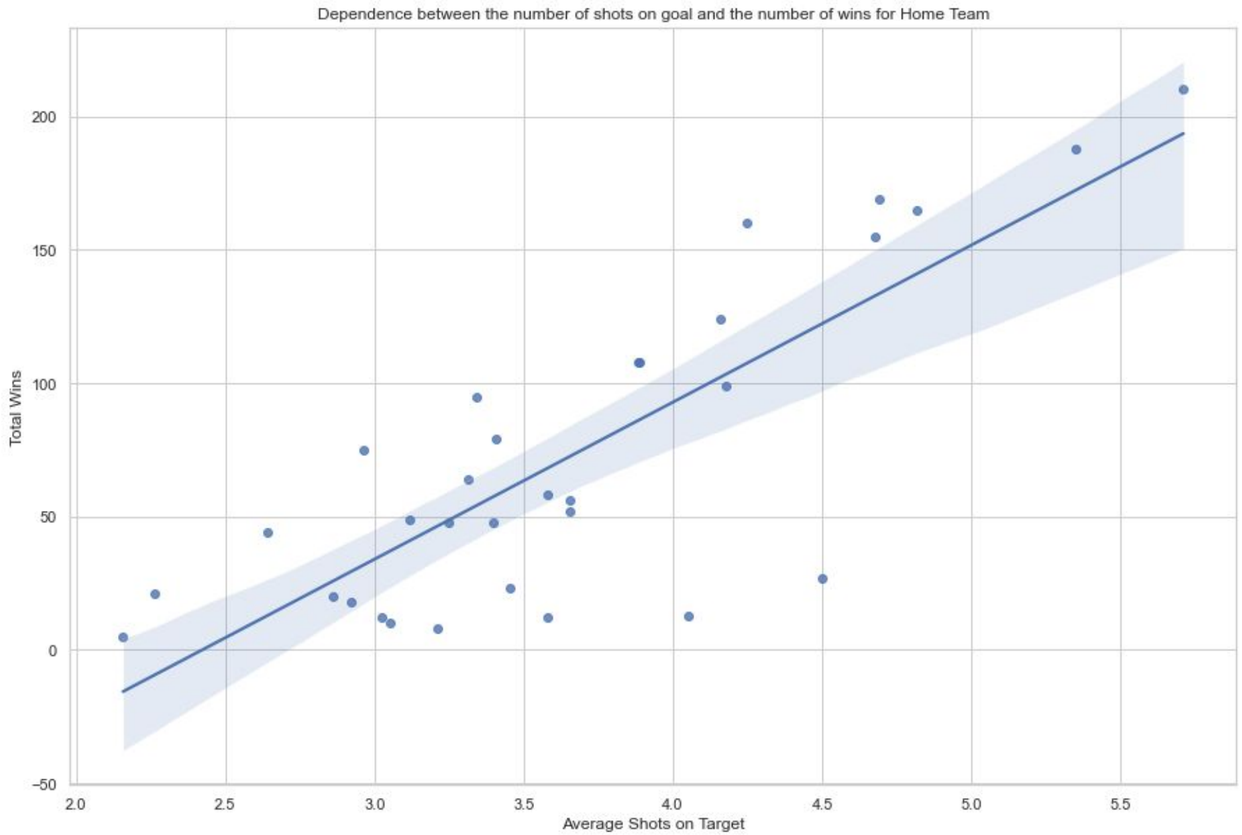
# Выводы:

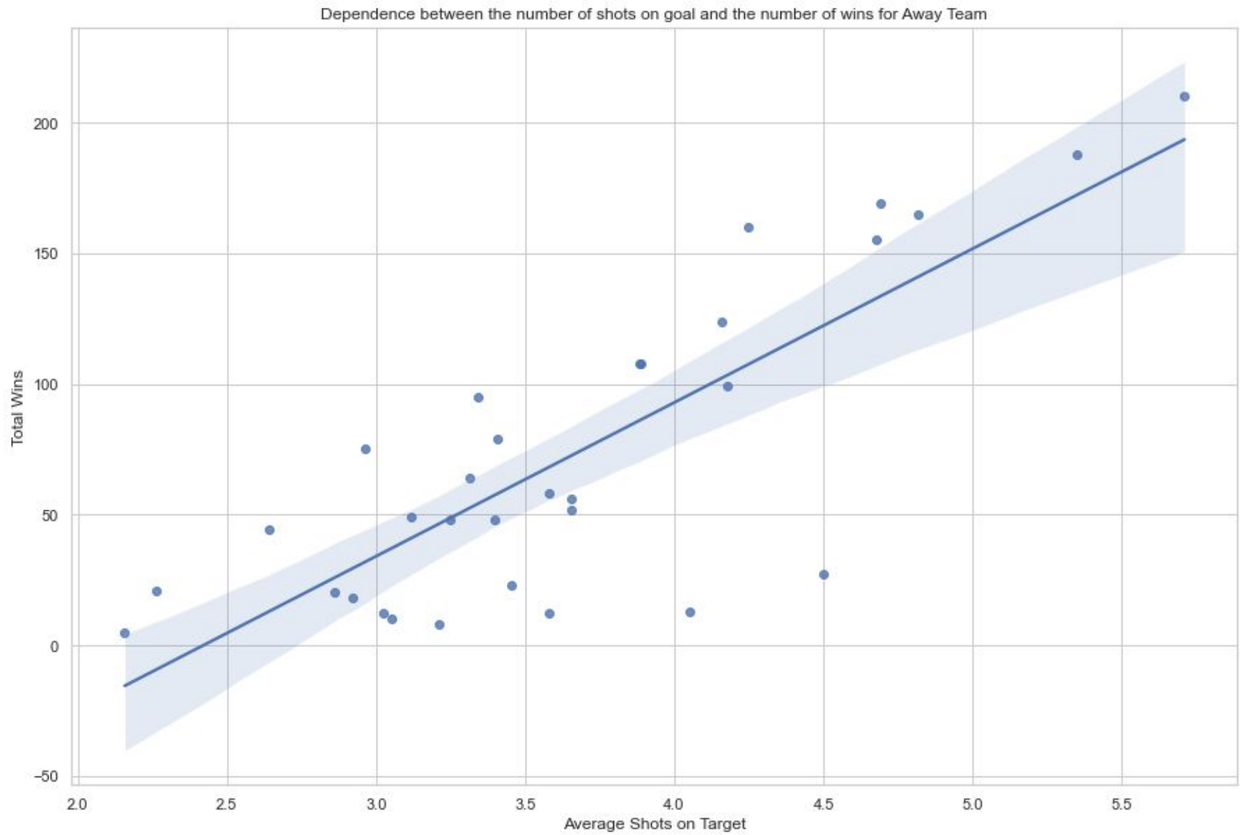
- Точность модели составляет примерно 57.6%. Учитывая, что точность бинарного классификатора составляет примерно 50%, модель логистической регрессии незначительно превосходит бинарный классификатор.

- Коэффициенты логистической регрессии: -0.231. Отрицательный коэффициент указывает на то, что с увеличением количества карточек вероятность победы команды уменьшается. Однако коэффициент достаточно мал, что может говорить о слабом влиянии количества карточек на исход матча.

- В целом, результаты анализа говорят о том, что влияние количества карточек на исход матча есть, но оно достаточно слабое.

# Команды, которые наносят большее количество ударов в створ ворот чаще выигрывают



# Выводы:

- С учетом полученного коэффициента корреляции Пирсона, равного 0.88 для домашних команд и 0.81 для гостевых команд, можно заключить, что существует сильная положительная зависимость между количеством ударов в створ ворот и количеством побед.

- Это значение подтверждает нашу гипотезу о том, что команды, которые наносят большее количество ударов в створ ворот, чаще выигрывают.

- Тем не менее, стоит учитывать, что корреляция не подразумевает причинно-следственную связь, и наличие других факторов может влиять на эту зависимость.

# Выводы ко всей работе.

В ходе анализа данных были проверены и подтверждены следующие гипотезы:

- Команды, играющие на домашнем стадионе, действительно побеждают чаще, чем в гостях.
- Ожидаемое количество голов (xG) средне положительно коррелирует с фактическим количеством голов, что указывает на то, что команды с более высоким показателем xG, в целом, забивают больше голов.
- Команды, которые ведут после первого тайма, чаще выигрывают матчи (доля таких матчей составляет примерно 77.1%).
- Команды, наносящие большее количество ударов в створ ворот, чаще выигрывают матчи.

- Гипотеза о том, что команды, зарабатывающие больше угловых ударов, чаще выигрывают или набирают больше очков, не подтвердилась.
- Также было выявлено, что влияние количества карточек на исход матча есть, но оно достаточно слабое.

# Пути развития и улучшения работы.

- Включение данных из других лиг, сезонов и турниров может помочь улучшить обобщающую способность модели и повысить точность прогнозов.

- Исследование и включение дополнительных переменных, таких как статистика игроков, состояние погоды, местоположение стадиона, состояние газона и т.д., может помочь улучшить точность модели и выявить новые факторы, влияющие на исход матча.

- Разработка моделей машинного обучения, таких как логистическая регрессия, случайный лес или градиентный бустинг, может повысить точность прогнозов исходов матчей, а также помочь выявить более сложные зависимости между переменными.

- Совмещение данных о матчах с данными о ставках может помочь лучше понять, какие факторы учитываются на рынке ставок и использовать эту информацию для определения наиболее выгодных ставок.


"That's all Folks!"