Отчёт по заданию 3 «Классификация методом kNN»

Выполнил студент 4 курса 7.1 группы Путин Павел

Вариант 11 «2004 New Car and Truck Data»

1. Описание данных:

Приведены технические характеристики 428 новых автомобилей, выпущенных в 2004 году. Зафиксированные параметры включают в себя цену, размеры автомобиля и его топливную экономичность.

Среди них 60 автомобилей – внедорожники (14%).

|  |  |
| --- | --- |
| Колонка | Имя параметра |
| 1-45 | Название транспортного средства |
| 47 | Спортивный автомобиль? (1=да, 0=нет) |
| 49 | Спортивный внедорожник? (1=да, 0=нет) |
| 51 | Универсал? (1=да, 0=нет) |
| 53 | Минивэн? (1=да, 0=нет) |
| 55 | Пикап? (1=да, 0=нет) |
| 57 | Полный привод? (1=да, 0=нет) |
| 59 | Задний привод? (1=да, 0=нет) |
| 61-66 | Рекомендованная розничная цена (в долларах США) |
| 68-73 | Дилерская стоимость (в долларах США) |
| 75-77 | Объем двигателя (л) |
| 79-80 | Количество цилиндров (=-1 для роторного двигателя) |
| 82-84 | Лошадиных силы |
| 86-87 | Миль на галлон по городу |
| 89-90 | Миль на галлон по шоссе |
| 92-95 | Вес (фунтов) |
| 97-99 | Колесная база (дюймов) |
| 101-103 | Длина (дюймов) |
| 105-106 | Ширина (дюймы) |

1. Решение
   1. Выделение обучающей и текстовой выборки

Обучающая выборка состоит из 84 объектов (17,9% от общего числа), среди них 12 внедорожников (14%) и 72 других автомобиля (86%) (см рисунок Рисунок 1). Обучение проходило по 6 критериям: наличие полного привода, объём двигателя, количество цилиндров, количество лошадиных сил, вес, колёсная база.

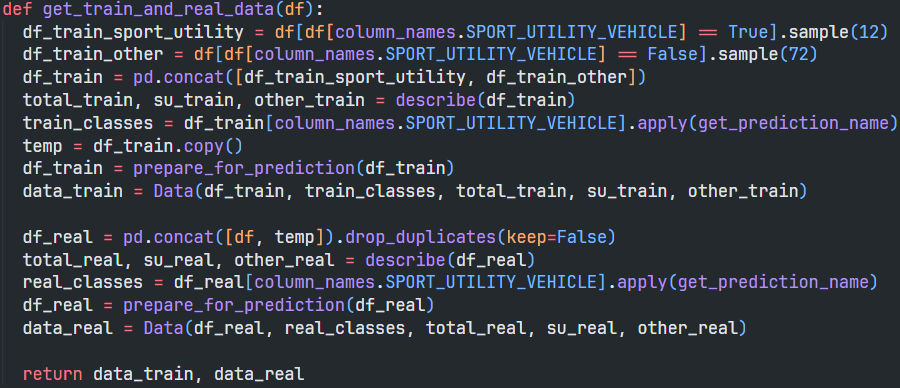


Рисунок - Метод выделения обучающей и тестовой выборки и контрольных значений классов

* 1. Определение наилучшего значения k

Для определения наилучшего значения k было выполнено по 50 запусков обучения модели на 50 разных обучающих данных (см рисунок Рисунок 2):

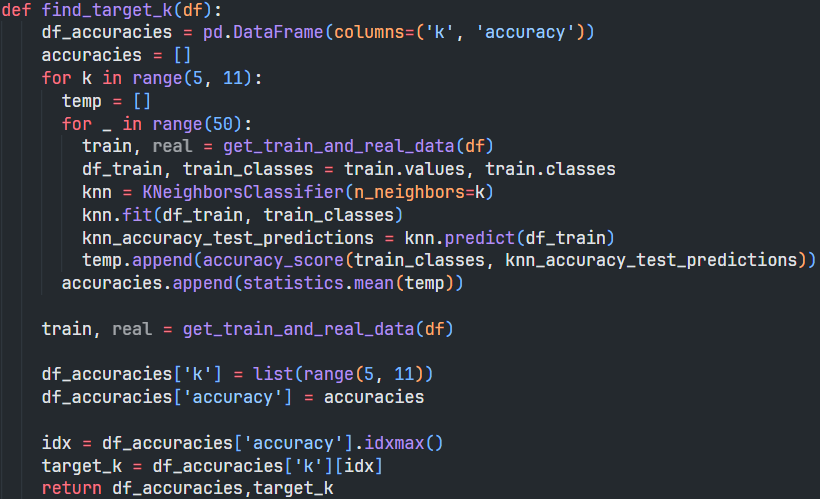


Рисунок - Определение оптимального значения k

На основе максимального среднего значения выбирается лучшее k. В силу случайности выборок, k может принимать значения 5 (чаще) или 6 (см рисунок Рисунок 3).

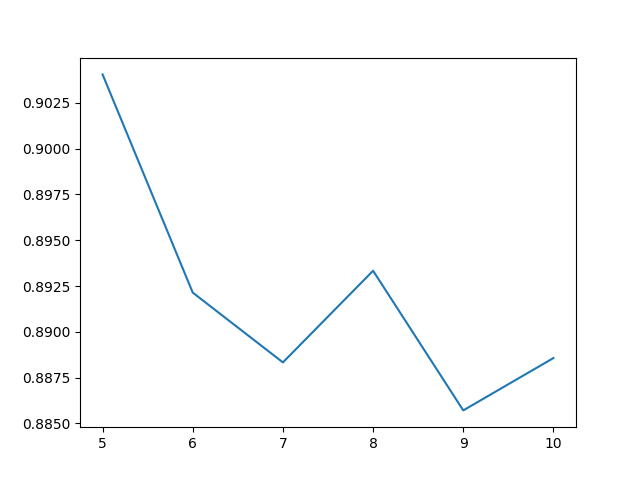


Рисунок - Зависимость точности классификатора от значения k

* 1. Оценка качества прогноза на тестовой выборке с помощью таблицы сопряженности

На основе зависимости количества автомобилей от типа и одного из параметров, участвовавших в обучении классификатора, была выявлена явная зависимость типа автомобиля от веса, наличия полного привода и колёсной базы (см рисунки Рисунок 4 - Рисунок 9).

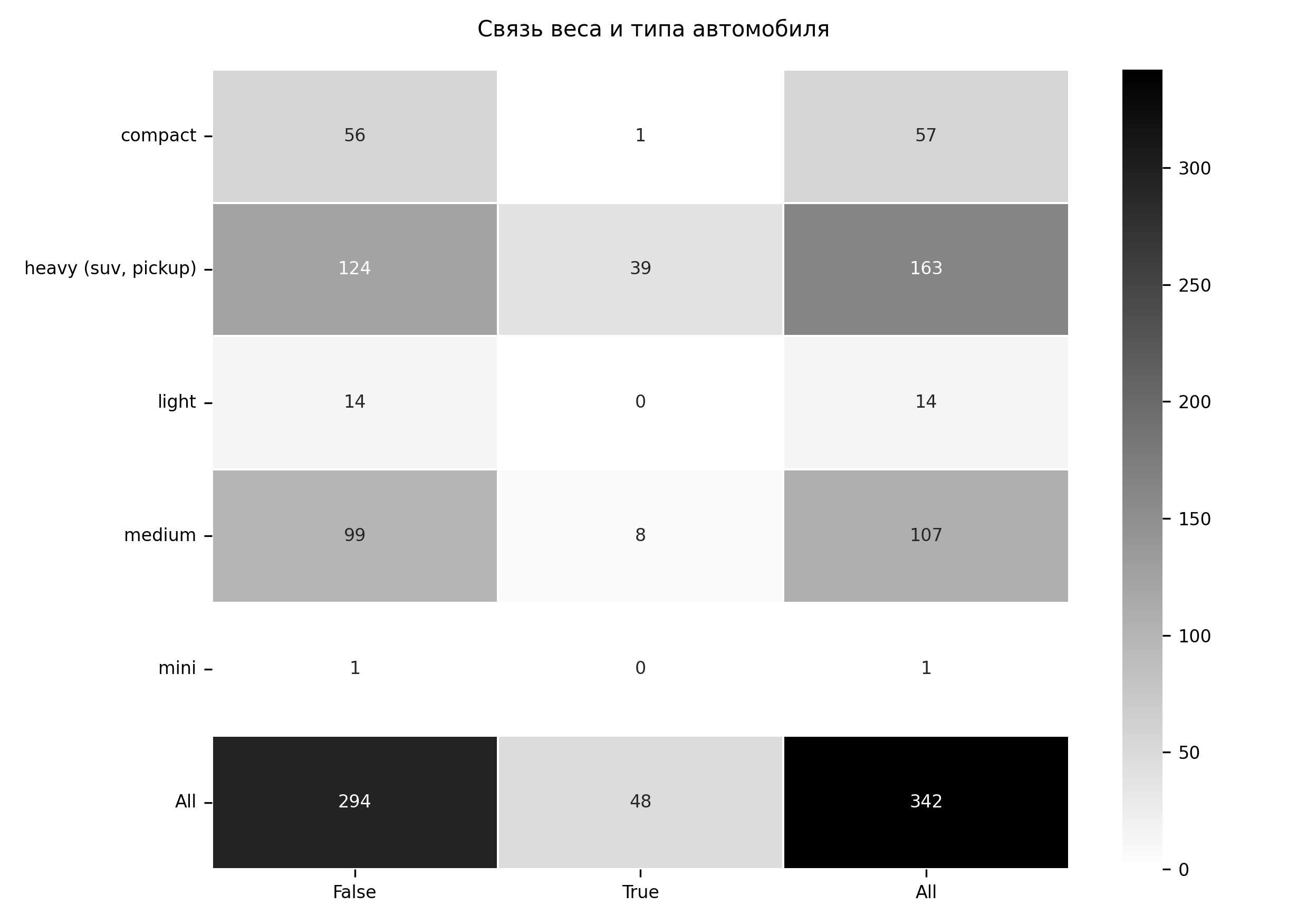


Рисунок - Связь веса и типа автомобиля

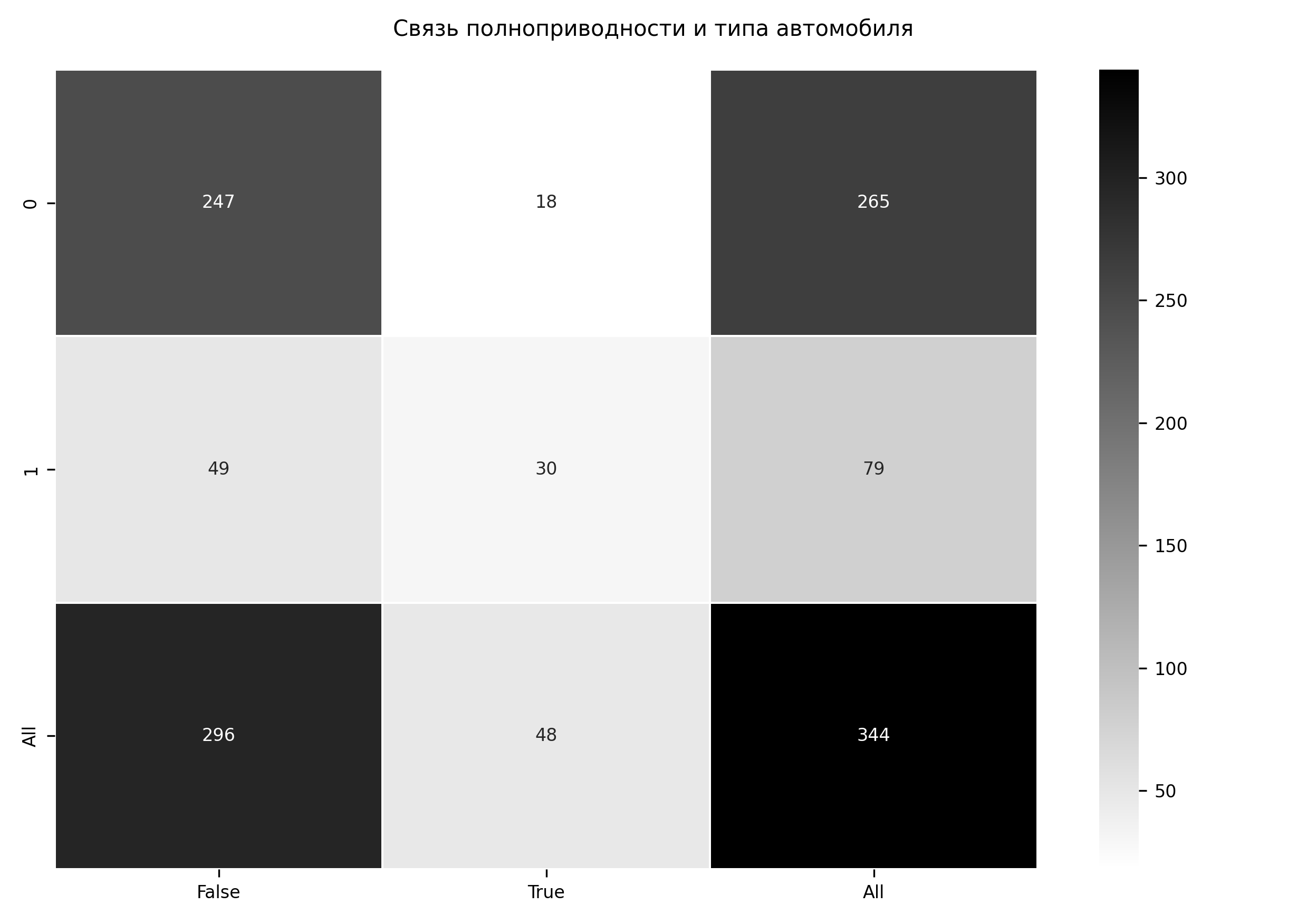


Рисунок - Связь наличия полного привода и типа автомобиля

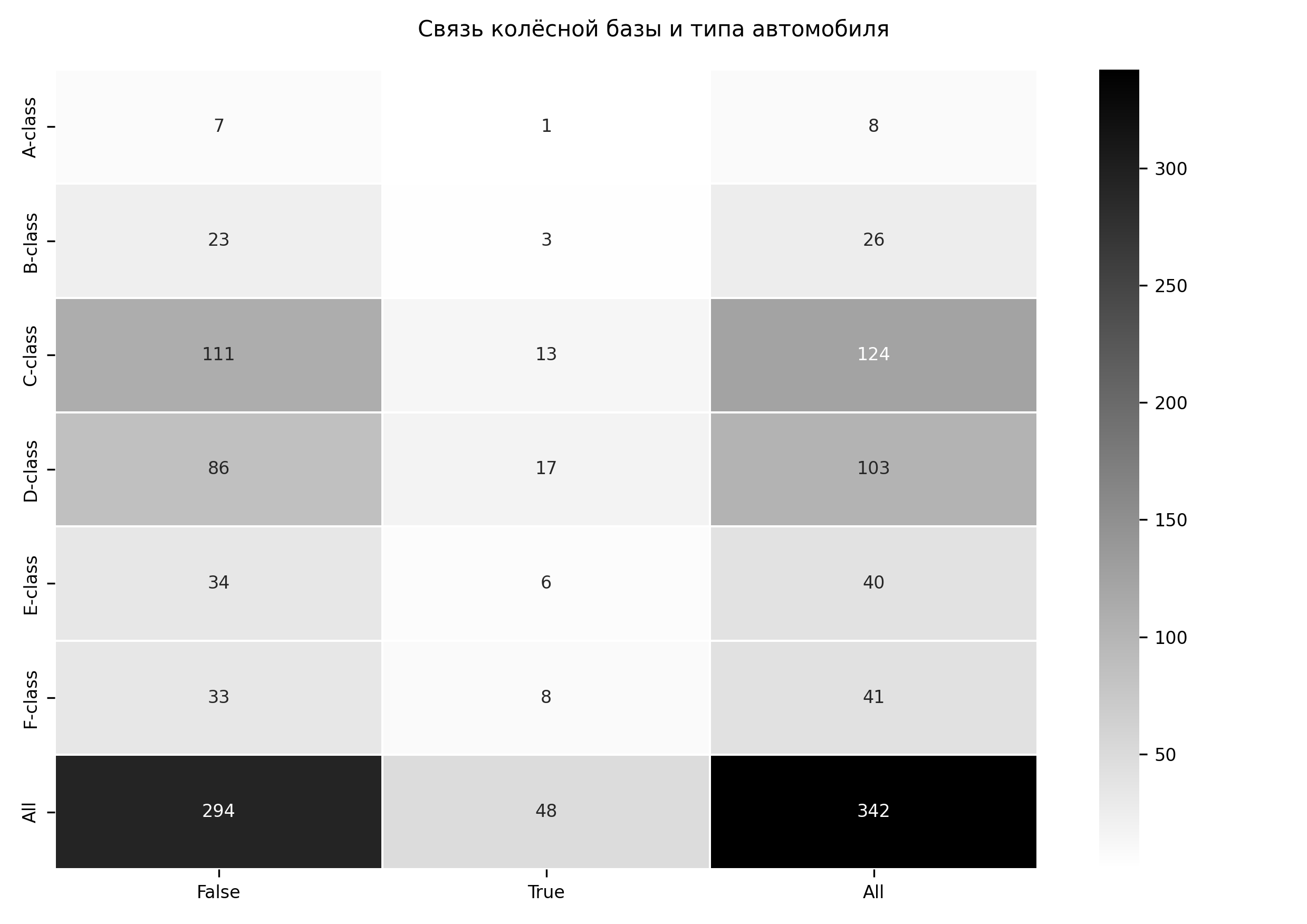


Рисунок - Связь колёсной базы и типа автомобиля

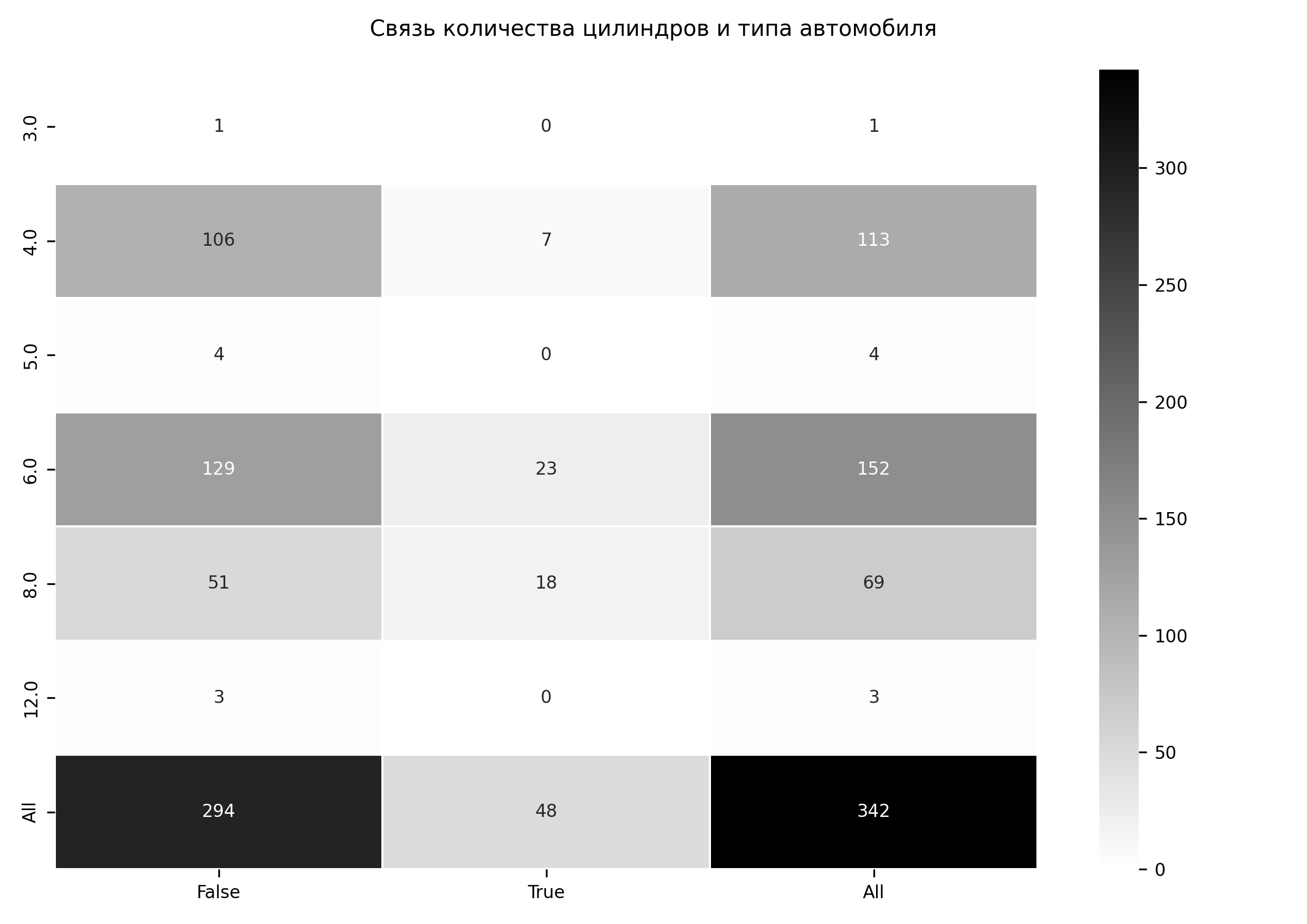


Рисунок - Связь количества цилиндров и типа автомобиля

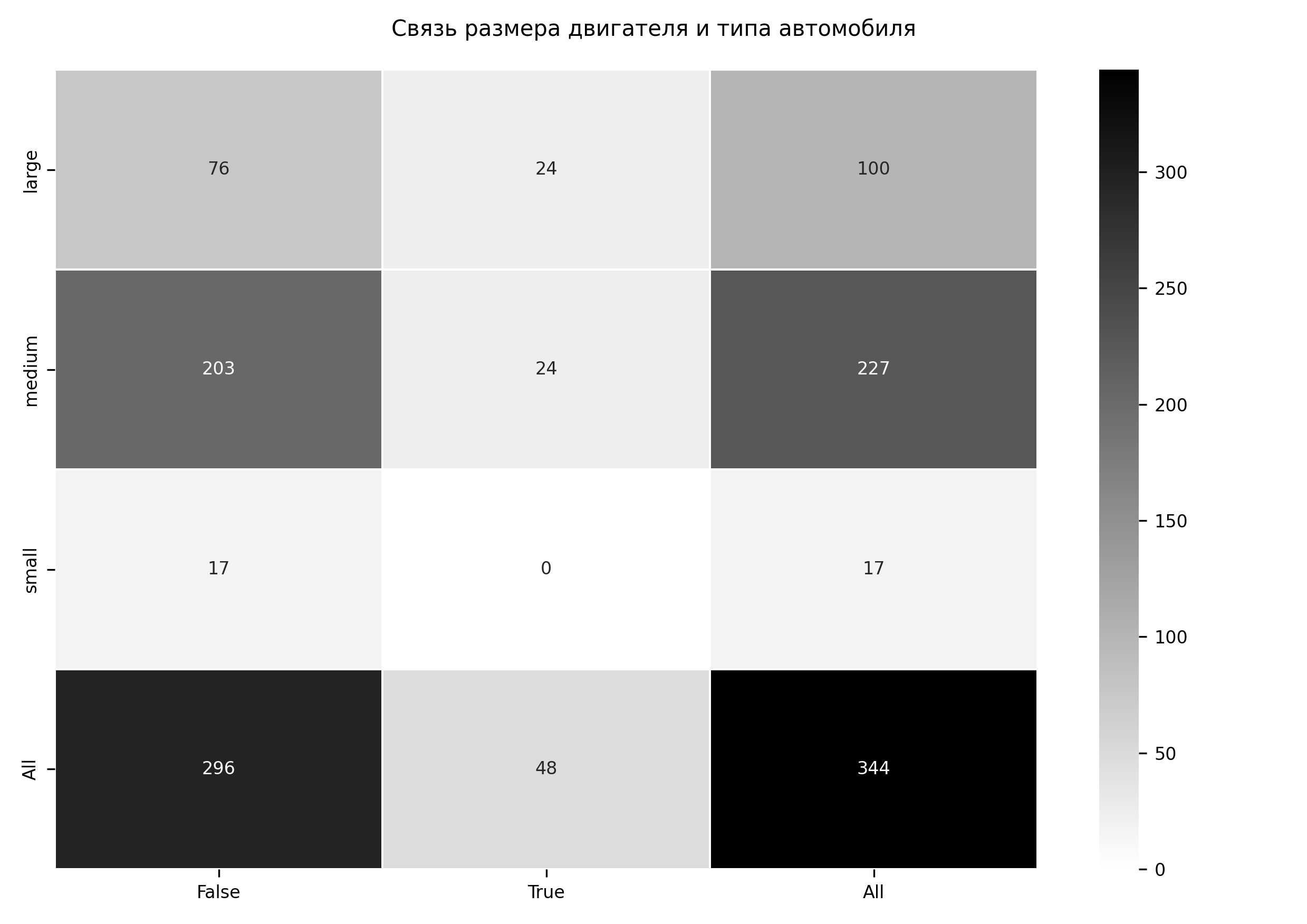


Рисунок - Связь объёма двигателя и типа автомобиля

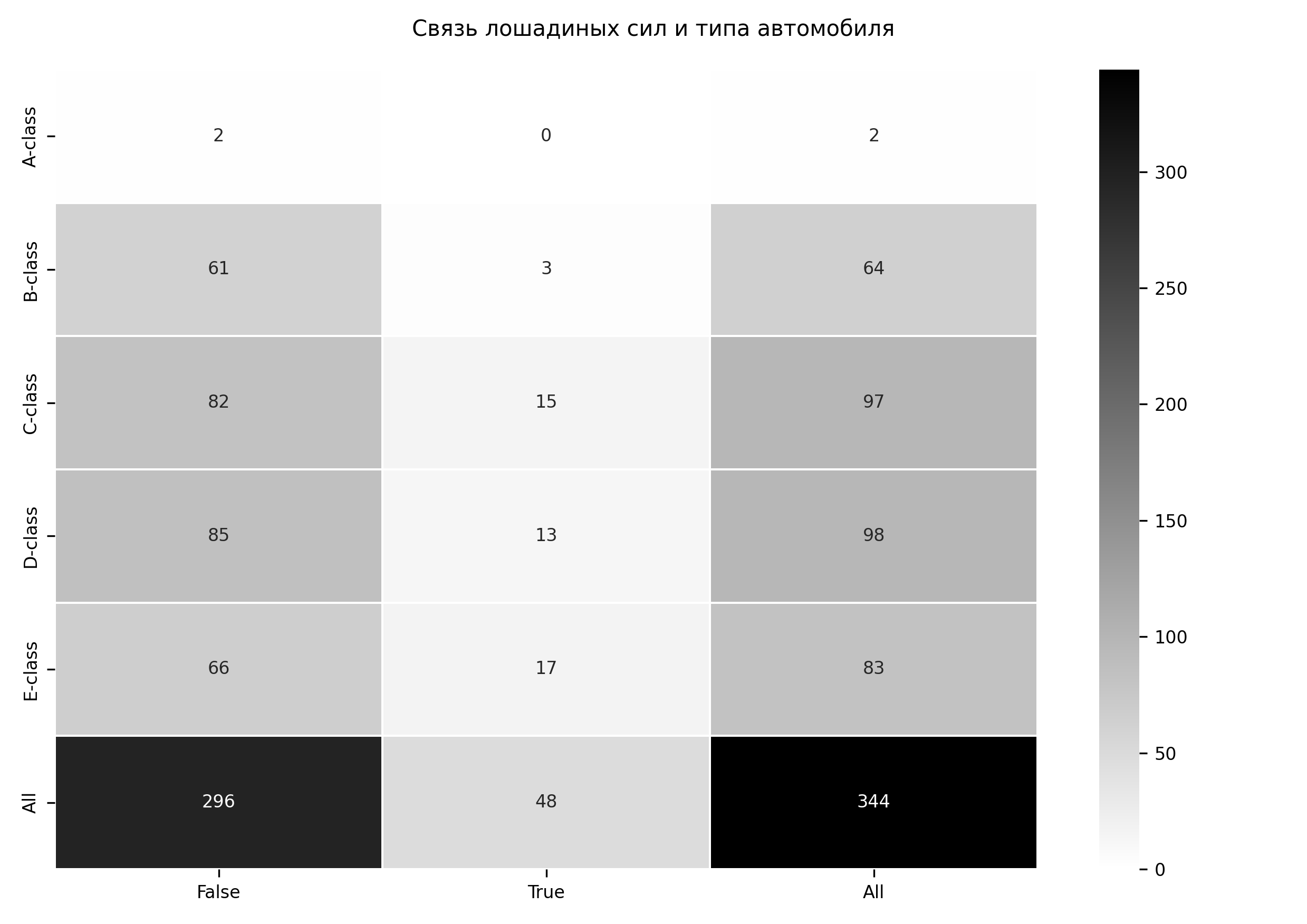


Рисунок - Связь количества лошадиных сил и типа автомобиля

1. Выводы

В итоге, был получен следующий результат точности классификации (k равен 5):

|  |  |
| --- | --- |
| Набор данных | Точность классификатора |
| Обучающие | 89,29% |
| Тестовые | 86,92% |