Исследование данных национального виктимизационного опроса

***Александр Дьяконов*, https://dyakonov.org/ag/**

В этом техническом отчёте представлены визуализации и сводные таблицы, которые сделаны с помощью данных, собранных Институтом проблем правоприменения (ИПП) при ЕУ СПб в ходе опроса населения.

Кроме настоящего технического отчёта автором подготовлены:

- тест, в котором в игровой форме задаются вопросы о результатах опроса; выложен в общий доступ по адресу
  https://forms.gle/QRcF2nt3D25iMmez7
  тест с ответами и пояснениями в pdf-формате выложен здесь:
  https://github.com/Dyakonov/visualization/blob/master/victimTEST_byDyakonov.pdf
- библиотека по визуализации результатов опроса для языка программирования Python 3 и учебный ноутбук (код с разметкой), в котором получены все представленные ниже визуализации; выложены в общий доступ по адресу
  https://github.com/Dyakonov/visualization

В ближайшем будущем планируется также публикация результатов в блоге автора (если не будет возражений со стороны ИИП, объявившего конкурс визуализации данных виктимизационного опроса).

1. Описание данных

Данные представляют результаты телефонного опроса респондентов не моложе 18 лет по технологии CATI на основании простой случайной выборки телефонных номеров [1]. Опрашивались не зависимо от гражданства, но понятно, что подавляющее большинство респондентов была гражданами РФ, см. рис. 1.1.

Все данные сведены в таблицу, см. табл. 1.1 (использовался файл ***rcvs_dataset_2019-06-21.tab***), для работы с которой использовалась библиотека

Pandas[1] языка программирования Python. Стоки таблицы (их 16818 шт.) соответствуют респондентам, столбы (189) – вопросам. Столбцы названы специальными кодовыми именами, расшифровка которых прилагается в файле ***codebook.html***. Например, IVDur – продолжительность интервью, Q1 – пол, ID – универсальный идентификатор респондента и т.п. (почти о всех признаках мы поговорим ниже).

ID	IVDur	Q1	Q2	Q75	Q75_1N	Q76	Q76_1N	Q5_1T	Q66	Q14	Q1414	Q18	Q15	Q16
12646573	544	1	57	1	2.0	2.0	NaN	полтора года тому назад	1.0	NaN	1.0	2.0	1.0	8.0
12658422	1643	1	59	1	4.0	1.0	2.0	открытый грабеж-выхватили сумку с документами ...	2.0	4.0	NaN	1.0	1.0	1.0
12660336	676	1	35	1	1.0	2.0	NaN	покупка на авито ,мы отправили деньги ,но нам...	1.0	NaN	14.0	2.0	1.0	12.0
12664831	510	1	22	1	1.0	2.0	NaN	украли телефон	2.0	1.0	NaN	2.0	1.0	6.0
12666214	1107	1	79	1	3.0	2.0	NaN	мошенничество,говорят и звонят я твой сын и до...	1.0	NaN	1.0	2.0	1.0	8.0

Табл. 1.1. Левый верхний угол таблицы с данными в формате библиотеки Pandas.

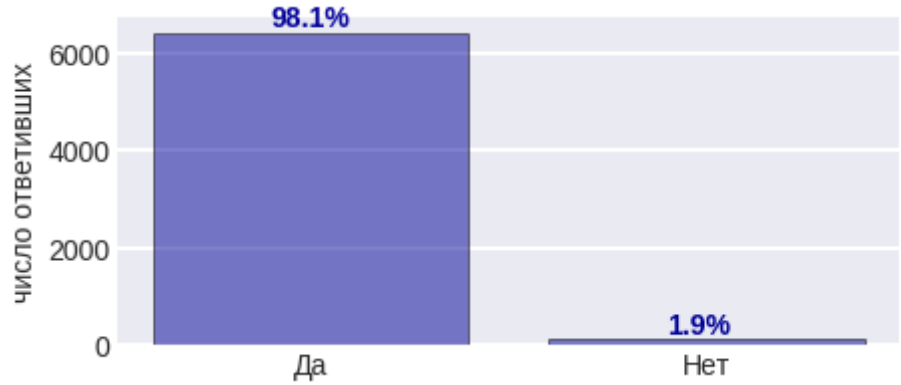


Рис. 1.1. У Вас есть гражданство Российской Федерации? [Да - 6409, Нет - 124].

Вопросы делятся на «анкетные» (пол, возраст, социо-демографический профиль), основной (были ли жертвой преступления – точная формулировка ниже) и уточняющие вопросы «о преступлениях» (где, когда и т.п.) Вопросы были составлены специальным образом профессионалами, учитывая многие особенности, например специфику русского языка [1]. В этом отчёте некоторые вопросы немного перефразированы (для краткости и красоты – чтобы умещались на изображениях).

[1] https://pandas.pydata.org/

2. Длительность интервью

Сначала посмотрим, сколько по времени проходили интервью. На рис. 2.1 показано распределение их длительностей[2]. Как показано на рис. 2.2, длительность связана с числом вопросов, на которые пришлось ответить респонденту. Число ответов примерно соответствует числу заполненных полей в таблице (там есть поля, которые заполняются автоматически, например иденификационный номер ID). Если человек не был свидетелем / участником преступления, то опрос длился в среднем 2 минуты, иначе ему приходилось ответить на уточняющие вопросы и опрос длился в среднем 10 минут.

Рис. 2.1. Плотность и гистограмма распределений продолжительности интервью.

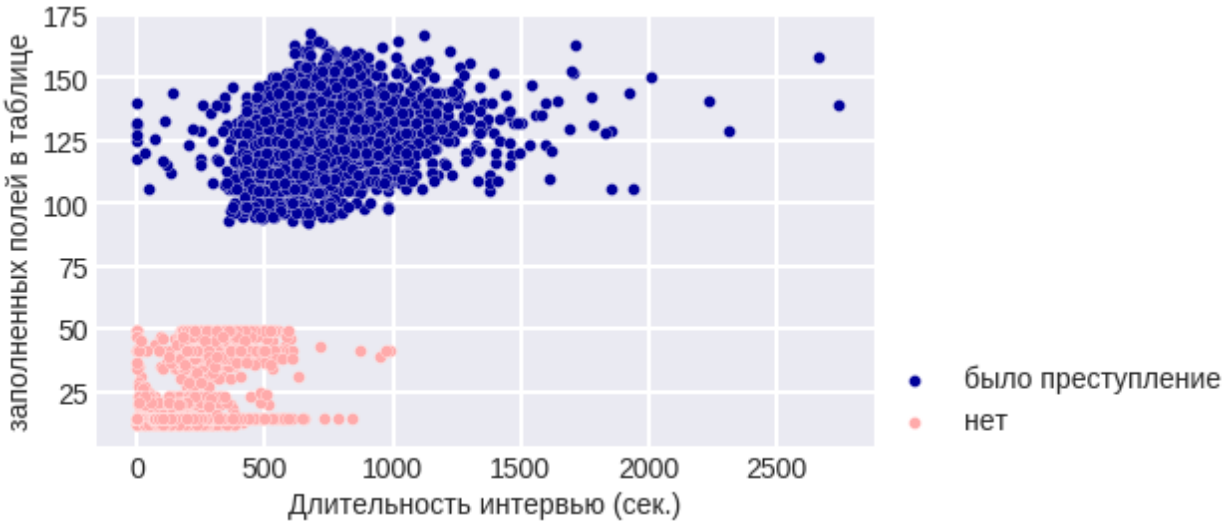


Рис.2.2. Диаграмма рассеивания по признакам: длительность и число заполненных полей.

Также на рис. 2.2 можно заметить странные выбросы. Во-первых, очень длинные интервью: больше 35 минут. Ниже ID этих интервью и описания последних преступлений из них.

[2] В техническом отчёте предполагается, что читатель знаком с понятиями «плотность распределения» и «гистограмма».

13368942 попытка изнасилования - выбросили с 4 Этажа
13376637 большая сумма денег была, решили изъять, вмешалось ФСБ, одному дали 18 другому 12 особого
13406546 Вытаскал к себе трубы (стека стояла). Которые я покупала сама.
14283011 клевета,оскорбление связано с вымогательством и захватом участка

Также подозрительно короткие интервью 0-1 сек, ниже ID этих интервью и описания последних преступлений из них.

13265395 заказали товар пришло не качественный. пытались связаться по телефону и смс не отвечали. месяц назад
13303505 кража кошелька из сумке при посадке на автобус
13378559 дали цыплят, половина подохли. Сказала, что буду платить только живых, угрожали, что спалят хату
13928814 угрозы
14163079 не выдали зарплату весной 2017 года
14301670 Избиение

Наличие подобных значений (продолжительность интервью 0 сек) показывает, что не все поля заполнены корректно, дальше мы ещё столкнёмся с этим. Среди описаний преступлений, например, встречаются такие:

13536071 htrkfvf ,skf jlyf? f jrfpsdftncz gbhfvblf abyfycjdfz

Понятно, что в данном случае при внесении информации забыли переключить раскладку и правильное описание «*реклама была одна, а оказывается пирамида финансовая*».

3. Основной вопрос

Основной вопрос интервью звучал так: **Q75 *«Вспомните, пожалуйста, было ли такое, что вас обокрали, вас побили, вам угрожали, вы стали жертвой насилия, мошенничества или других преступлений в России за последние 5 лет?».*** Ниже в табл. 3.1 показана статистика ответов на него

	ответов	процентов
Нет	13776	81.9%
Да	3001	17.8%
Затрудняюсь ответить / не помню	41	0.2%

Табл. 3.1. Статистика ответов на основной вопрос.

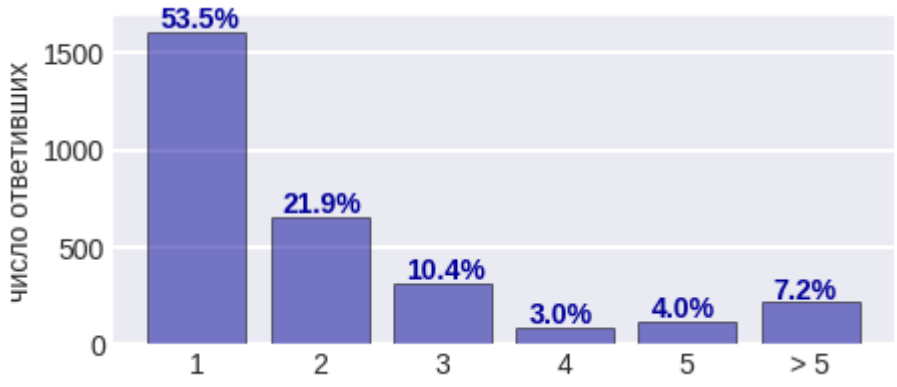
Интересно, что 18% респондентов не ответили на этот вопрос «Нет». Ниже в табл. 3.2 показана статистика с разбиением по полу. Проценты здесь от суммы всех чисел в таблице. Сначала автору подобная цифра показалась гигантской, но в процессе анализа данных выяснилось, что «преступление» трактуется довольно широко (для логики автора), в частности, преступлением является

СМС-мошенничество, когда Вам приходит СМС с просьбой перевести деньги (и многие отвечали на вопросы, указывая это как последнее преступление). После анализа описания преступлений, 18% кажется уже заниженным показателем преступности.

	Да	Нет	Затрудняюсь ответить / не помню
Ж	1618 (9.6%)	7663 (45.6%)	13 (0.1%)
М	1383 (8.2%)	6113 (36.3%)	28 (0.2%)

Табл. 3.2. Статистика ответов на основной вопрос с разбиением по полу.

Есть также, уточняющие вопросы, например, **Q75_1N** ***«Сколько всего подобных случаев произошло с Вами за последние 5 лет?»***. При анализе ответов на них, в данных видные некорректные значения. Так на этот вопрос в столбце присутствуют значения: «-9223372036854775808» (явный выброс), «2013» (это скорее год), «1000», «3000», «1830» (слишком большие числа) и т.п. Для упрощения все записи с числом преступлений больше 5 было решено объединить в одну категорию, см. рис. 3.1.



	ответов	процентов
1	1606	53.5%
2	658	21.9%
3	311	10.4%
4	90	3.0%
5	120	4.0%
> 5	216	7.2%

Рис. 3.1. Статистика ответов на вопрос о числе преступлений за последние 5 лет.

Ниже представлена сводная табл. 3.3: как отвечали на пару вопросов «по числу преступлений за последние 5 лет» и «были такие преступления в последнем году», а также табл. 3.4, уточняющая число преступлений в прошлом году.

	1	2	3	4	5	> 5
Да	481 (16.0%)	294 (9.8%)	185 (6.2%)	54 (1.8%)	91 (3.0%)	183 (6.1%)
Нет	1121 (37.4%)	362 (12.1%)	125 (4.2%)	33 (1.1%)	27 (0.9%)	31 (1.0%)
Затрудняюсь ответить / не помню	4 (0.1%)	2 (0.1%)	1 (0.0%)	3 (0.1%)	2 (0.1%)	2 (0.1%)

Табл. 3.3. Сводная таблица: число преступлений за 5 лет (по горизонтали) и наличие их в прошлом году (по вертикали).

	1	2	3	4	5	> 5
1	466 (36.2%)	207 (16.1%)	121 (9.4%)	18 (1.4%)	32 (2.5%)	30 (2.3%)
2	5 (0.4%)	77 (6.0%)	37 (2.9%)	23 (1.8%)	36 (2.8%)	30 (2.3%)
3	6 (0.5%)	6 (0.5%)	20 (1.6%)	3 (0.2%)	15 (1.2%)	37 (2.9%)
4	0 (0.0%)	0 (0.0%)	1 (0.1%)	9 (0.7%)	1 (0.1%)	17 (1.3%)
5	1 (0.1%)	0 (0.0%)	6 (0.5%)	0 (0.0%)	6 (0.5%)	22 (1.7%)
> 5	3 (0.2%)	4 (0.3%)	0 (0.0%)	1 (0.1%)	1 (0.1%)	47 (3.6%)

Табл. 3.4. Сводная таблица: число преступлений за 5 лет (по горизонтали) и число их в прошлом году (по вертикали).

4. Машинное обучение и анкетные данные

Выше отмечалось, что в анкете часть вопросов/ответов не была связана с преступлениями, а описывала респондента (пол, возраст, образование и т.п.) Было решено построить модель машинного обучения, которая по описанию человека определяет, сталкивался ли он с преступлением. Ясно, что есть некоторые возражения против осмысленности такой модели:

- данных не так много (если респондент был жертвой преступления, то проходил полный опрос – таких было 3001 человек, из остальных только у 3719 спрашивали анкетные данные),
- постановка задачи не совсем корректна, т.к. если человек становился жертвой преступления, то это был совершившийся факт (в котором, впрочем, есть доля случайности), а если нет, то это могло случиться с ним вскоре после интервью, т.е. его анкетные описания – это не описания человека, с которым точно ничего не случится.

Отметим, что во-первых, наша модель не будет использоваться на практике, а потребуется нам для нахождения закономерностей в данных (которые можно при желании верифицировать другими методами). Во-вторых, проблема малой для машинного обучения выборки решается использованием несложных стабильных моделей. В данном случае применялись методы, основанные на ансамбле неглубоких решающих деревьев[3], они параллельно позволяют оценивать важность признаков. В-третьих, в подобной постановке традиционно решаются задачи в банковской отрасли, например, в скоринге (по описанию

[3] Библиотека https://lightgbm.readthedocs.io/en/latest/ для ЯП Python.

клиента определить, вернёт ли он кредит). Практика показывает, что модели машинного обучения в подобных постановках вполне разумны.

Замечание. Интересно, что тот факт, что анкетные данные заполняли не все, автор выяснил с помощью машинного обучения. В начале работы с данными, автор попытался не читать описания, а понять смысл данных с помощью стандартных средств анализа и интерпретации данных. Удалось построить модель, классифицирующую респондентов на классы «было преступление» / «не было», довольно высокого качества, но при её анализе было обнаружено, что она существенно использует факт отсутствия ответов на определённые вопросы, т.е. правильно классифицирует тех, кому вопросы не задавали. Визуальный анализ их статистики показал, что все они не были свидетелями преступления.

Замечание. В реальности, данных для построения модели было даже меньше, чем описаний 3001 + 3719 респондентов. Во-первых, использовались только описания респондентов, которые чётко ответили на вопрос «были ли они жертвой преступления» (только «да» и «нет»). Во-вторых, некоторые вопросы не всем задавались, например вопрос Q80 «Сколько человек, считая Вас, живет с Вами вместе и ведет общее хозяйство?», т.е. кроме понятного ответа «затрудняюсь ответить» в таблице есть просто незаполненные поля. В итоге, для построения модели отобрано 5796 анкетных данных.

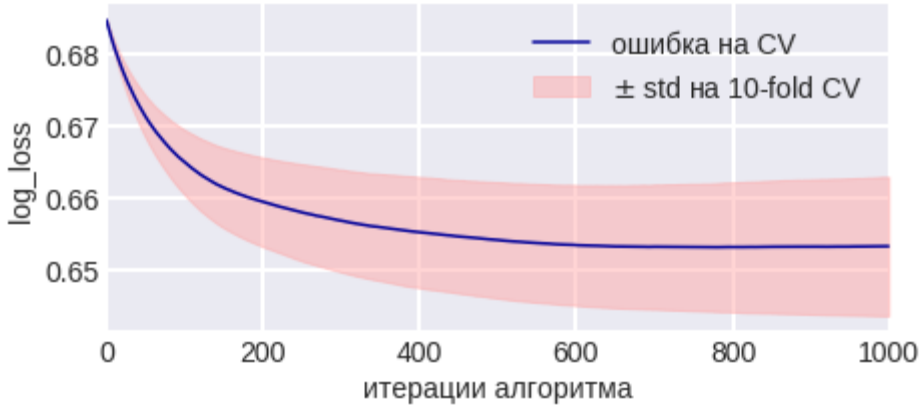


Рис.4.1. Логистическая ошибка при обучении алгоритма градиентного бустинга над деревьями

Опустим в отчёте эксперименты по подбору и настройки моделей. Были использованы признаки из рис. 4.2, а логистическая ошибка модели при

обучении показана на рис. 4.1[4]. В результате т.н. матрица ошибок (несоответствий)[5] модели выглядит так:

	$a = 0$	$a = 1$
$y = 0$	TN = 2493	FP = 773
$y = 1$	FN = 1607	TP = 923

Табл. 4.1. Матрица несоответствий, по горизонтали – ответы алгоритма, по вертикали – наличие преступления (1 – было преступление, 0 – не было).

Например, если модель показывает, что анкетные данные свидетельствует о том, что это жертва преступления (a=1), то в 923 / (923 + 773) доле случаев модель права (около 54%), если говорит, что это не жертва преступлений, то права в 2493 / (2493 + 1607) доле случаев (около 61%). В целом, качество довольно низкое. Однако, можно посмотреть на важность признаков, которые оценила модель (важность дана в условной шкале, в которой их выдаёт встроенный метод библиотеки lightgbm), см. рис. 4.2.

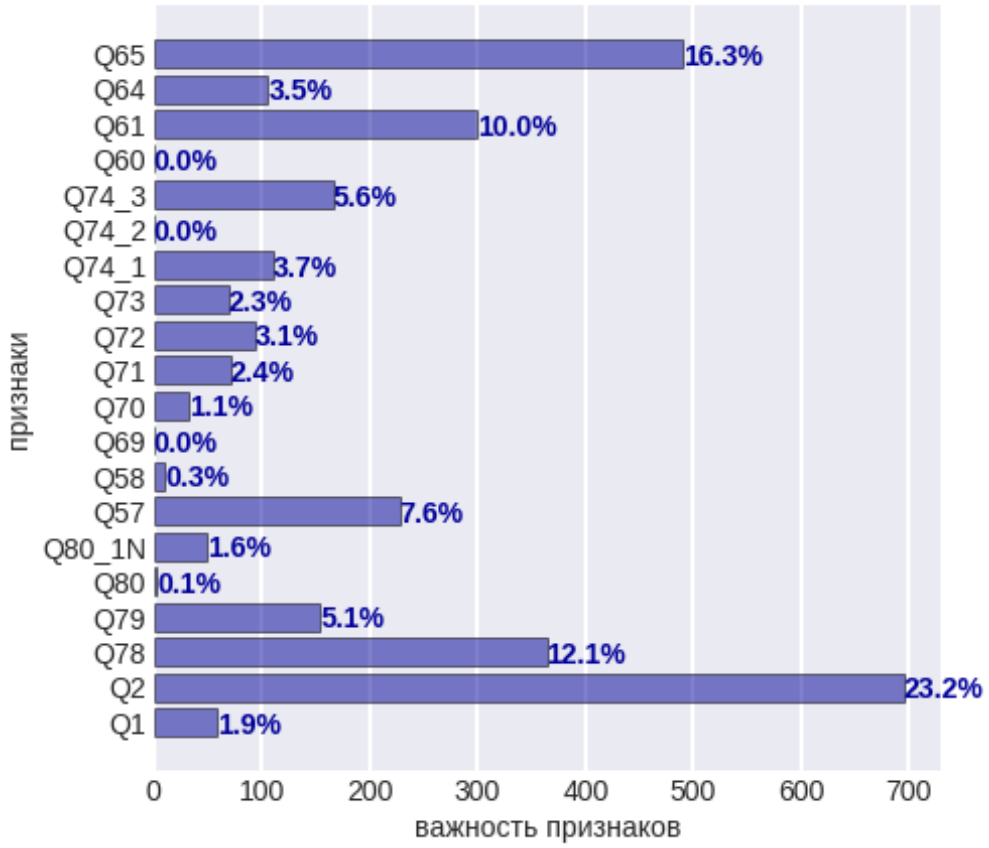


Рис. 4.2. Важности признаков, построенные моделью

[4] https://clck.ru/HNj5B

[5] https://clck.ru/HNj8F

На рис. 4.2 не показаны расшифровки признаков, поскольку в следующем разделе мы поговорим о самых важных отдельно.

5. Важные признаки в анкетных данных

Q2 Скажите, пожалуйста, сколько Вам полных лет?

Возраст оказался самым важным признаком с точки зрения модели из предыдущего раздела. На рис. 5.1 показано распределение возрастов респондентов. Отдельно изображено число мужчин и женщин каждого дискретного ответа «число полных лет». Отметим, однако, что признак «Пол» не считался моделью важным. Также отметим, что до 40 лет (включительно) среди респондентов больше мужчин (4230 против 4108), а после 40 существенно больше женщин (5186 против 3294). В [1] отмечено, что это согласуется со статистикой по РФ.

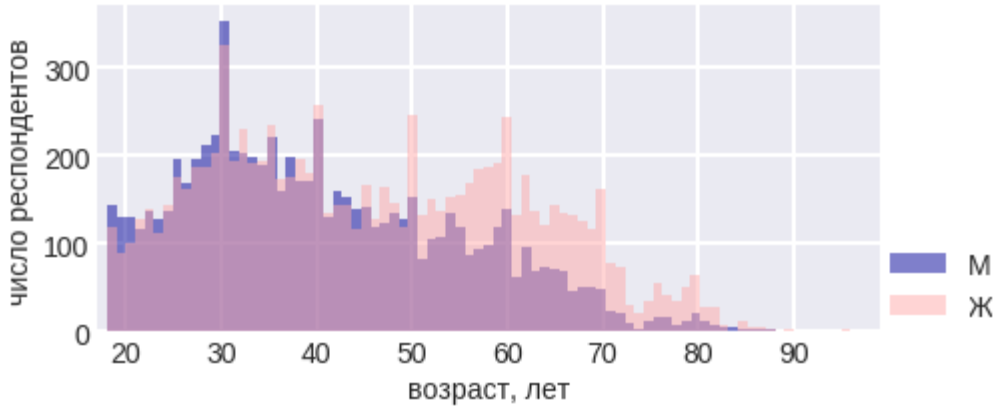


Рис.5.1. Распределение респондентов по возрасту и полу

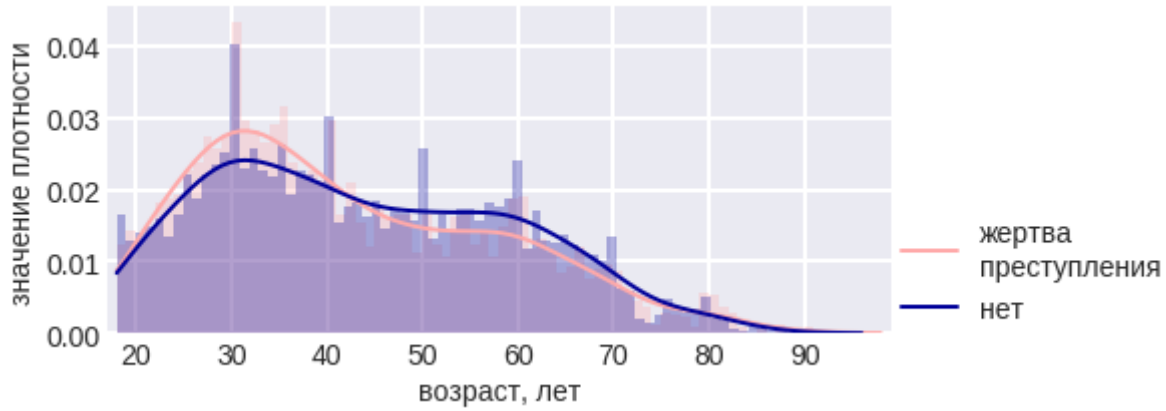


Рис.5.2. Распределение возрастов жертв преступлений и остальных респондентов

Всем респондентам не меньше 18 (по условиям проведения опроса), максимальный возраст – 98! Как всегда, в таких случаях «круглые числа» типа 30 лет, 40 лет и т.п. называют чаще (округляя свой возраст), см. рис. 5.1.

Важность возраста как признака в модели из раздела 4 объясняется тем фактом, что чем ниже возраст, тем больше шанс стать жертвой преступления (в статистике больше процент жертв), с поправкой на ограничение 18+ при опросе, см. рис. 5.2, а также табл. 5.4.

Q65 Привлекались ли Вы сами когда-нибудь к уголовной ответственности?

Здесь и далее в таблицах, начиная с 5.1, показано, сколько респондентов отвечали на основной вопрос определённым образом (Да, Нет, Затрудняюсь ответить / не помню – сокращено до «не знаю») и определённым образом на выбранный вопрос (достаточно важный с точки зрения модели). В правой части таблиц показано, в скольких процентах случаях при определённом ответе на выбранный вопрос человек был жертвой (процент считается только по респондентам, которые уверенно ответили на основной вопрос). **Важно**, что проценты в таблицах не надо интерпретировать как вероятность стать жертвой преступления. Напоминаем, что полный анкетный опрос проходили все жертвы преступлений и примерно столько же не-жертв, поэтому процент жертв в этой выборке близок к 50% (т.е. сильно завышен). Но проценты в таблицах можно сравнивать, что мы и будем делать (это ничему не противоречит)!

	да	нет	не знаю		процент
да	226	195	4	**да**	53.7%
нет	2775	3283	33	**нет**	45.8%

Табл. 5.1. Ответы на основной вопрос (были ли жертвой преступления – слева вверху) и Q65 «Привлекались ли Вы сами когда-нибудь к уголовной ответственности?» (левый столбец), справа – процент случаев, когда респондент становился жертвой при определённом ответе на вопрос Q65

Логично было предположить, что те, кто сам привлекался к уголовной ответственности, чаще сталкиваются с преступлением – это подтвердилось, причём разница в процентах заметна (здесь не будем отдельно оценивать статистическую значимость результатов). Интересно, что среди тех, кому задавался вопрос про уголовную ответственность 6.5% ответили, что привлекались, см. рис. 5.3. Даже если учесть только что полученный вывод, что таких респондентов больше среди жертв, а их чуть меньше половины среди тех, кому вопрос задавался, всё равно получается неожиданно (для автора) большой процент среди случайной выборки населения – больше 5.5%!

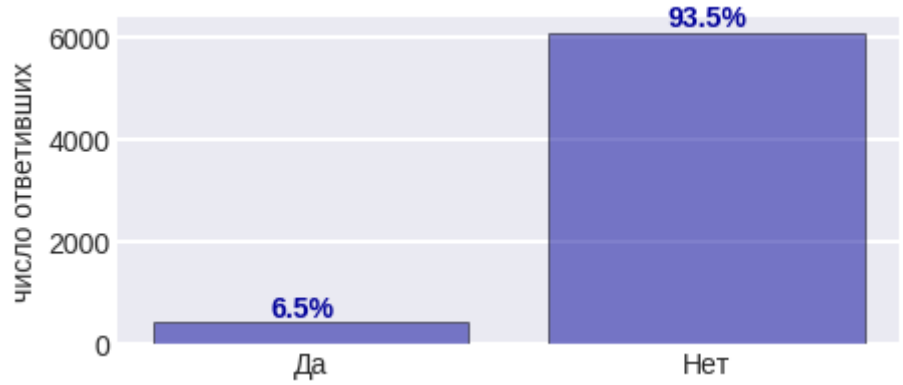


Рис. 5.3. Привлекались ли Вы сами когда-нибудь к уголовной ответственности? [Да - 425, Нет - 6091]

Q78 Скажите, Вы проживаете один или с кем-то?

	да	нет	не знаю	процент
один	509	573	6	47.0%
с кем-то	2481	3076	31	44.6%
затрудняюсь/не скажу	11	60	0	15.5%

Табл. 5.2. Ответы на основной вопрос (были ли жертвой преступления – слева вверху) и Q78 «Скажите, Вы проживаете один или с кем-то?» (левый столбец), справа – процент случаев, когда респондент становился жертвой при определённом ответе на вопрос Q78

Здесь мы чуть упростили на таблице варианты ответов; правильные: «Один», «С кем-то», «Затрудняюсь ответить, отказ говорить». Логично было бы предположить, что одинокие сталкиваются с преступлениями чаще, – статистика это подтверждает. Интересно, что те, кто не говорит о личной жизни, существенно реже сталкивались с преступлениями (15.5% против 44.6% и 47%), но этот процент вычислен на 71 анкете, поэтому точно не является статистически значимым, см. также рис. 5.4.

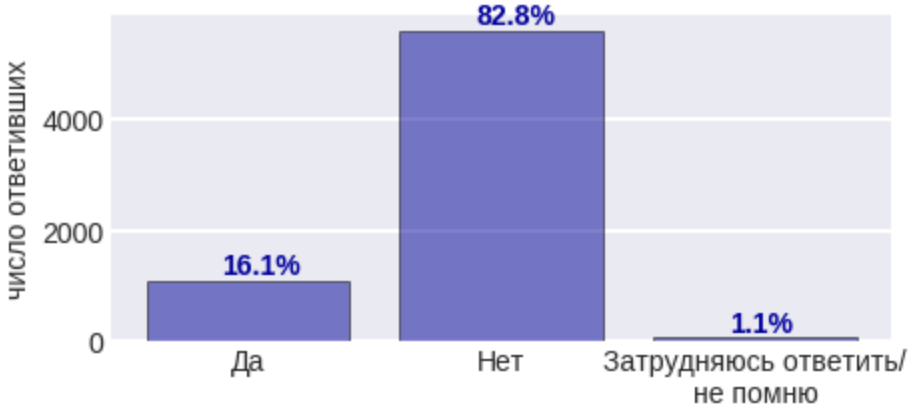


Рис.5.4. Вы проживаете один? ['Да - 1088', 'Нет - 5588', 'Затрудняюсь /не помню - 71']

Q61 Какое у Вас образование?

	да	нет	не знаю	процент
Полное среднее и ниже	543	813	7	40.0%
Среднее спец-ное/техническое или нач-ное профес-ное	1060	1348	14	44.0%
Высшее и незаконченное высшее	1398	1329	16	51.3%

Табл. 5.3. Ответы на основной вопрос (были ли жертвой преступления – слева вверху) и Q61 «Какое у Вас образование?» (левый столбец), справа – процент случаев, когда респондент становился жертвой при определённом ответе на вопрос Q61

Вот тут автора ждал небольшой сюрприз, см. табл. 5.3. Интуитивно, чем выше уровень образования, тем меньше шансов стать жертвой преступления. Статистика говорит об обратном, причём проценты существенно отличаются. Возникает гипотеза, что, возможно, это связано с возрастом. Понятно, что высшее образование получают позднее, на рис. 5.5 можно заметить (это лучше видно при большем масштабе и разрешении) сгусток точек в области до 20 лет с образованием не выше среднего, аналогичные сгустки есть для среднего и высшего образований (смещены левее).

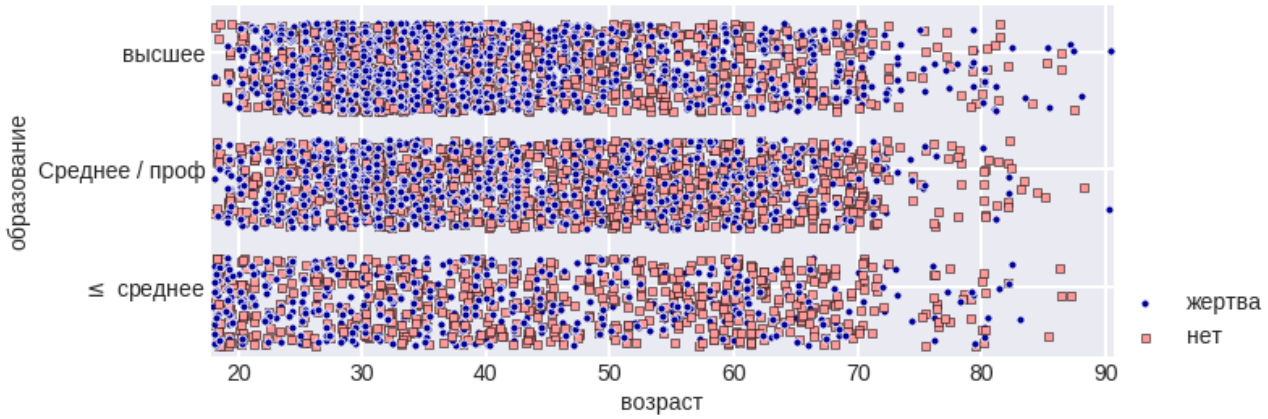


Рис. 5.5. Диаграмма рассеивания (скатерплот) по признакам «возраст» и «образование»

Однако, если людей разбить по категориям возраста, см. табл. 5.4, то во всех возрастных группах, чем выше уровень образования, тем больше процент жертв преступлений. Распределение ответов на вопрос об образовании показано на рис. 5.6.

	18-29	30-49	49-99
Полное среднее и ниже	47.6%	40.9%	34.1%
Среднее спец-ное/техническое или нач-ное профес-ное	49.0%	45.3%	40.4%
Высшее и незаконченное высшее	51.6%	53.4%	47.3%

Табл. 5.4. Процент случаев, когда респондент становился жертвой при определённом ответе на вопрос Q61, в разных возрастных группах.

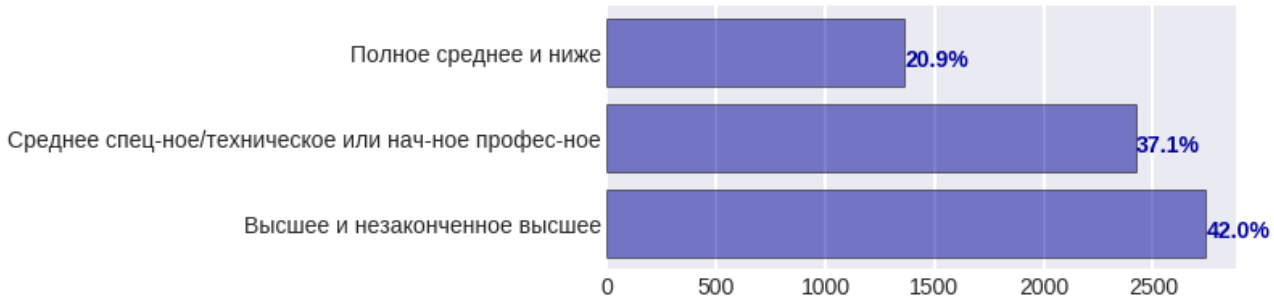


Рис. 5.6. Распределение ответов на вопрос «Какое у Вас образование?»

Q57 Как бы вы определили свой уровень дохода?

	да	нет	не знаю	вероятность
Едва сводим концы с концами, денег не хватает на продукты	294	297	5	49.7%
На продукты хватает, на одежду нет	800	969	13	45.2%
На продукты и одежду хватает, на технику и мебель нет	1040	1227	8	45.9%
На технику и мебель хватает, на большее денег нет	552	612	7	47.4%
Можем позволить автомобиль, но квартиру или дачу нет	137	159	3	46.3%
Можем позволить себе практически все: квартиру и т.д.	87	109	0	44.4%

Табл. 5.5. Ответы на основной вопрос (были ли жертвой преступления – слева вверху) и Q57 «Как бы вы определили свой уровень дохода?» (левый столбец), справа – процент случаев, когда респондент становился жертвой при определённом ответе на вопрос Q57

Интуитивно, чем выше доход, тем меньше процент жертв преступлений. Для высокого дохода – 44.4%, для низкого – 49.7%, но вот все промежуточные стадии, см. табл. 5.5, слабо различимы – от 45.2 до 47.4% (и в них уже нет монотонности процента от дохода). Скорее всего, это связано также с тем, что человеку сложно формализовать свой доход в таких категориях.

Остальные признаки оказались менее важными с точки зрения модели, поэтому просто приведём распределения ответов по ним.

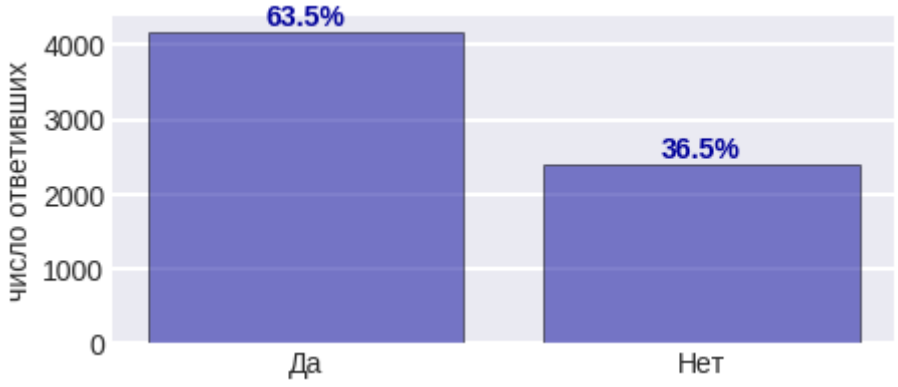


Рис. 5.7. Вы работаете в настоящее время? ['Да - 4174', 'Нет - 2395']

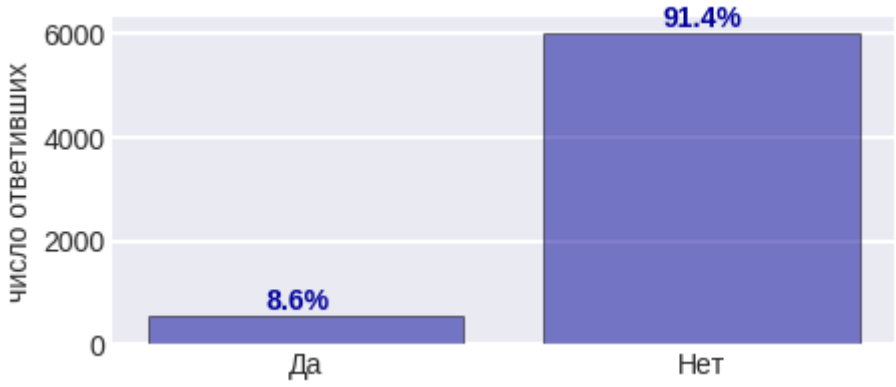


Рис. 5.8. Вы учитесь в настоящее время? ['Да - 564', 'Нет - 5999']

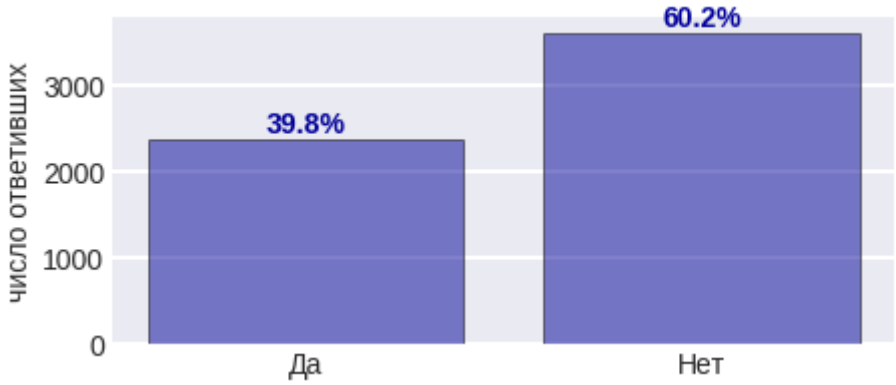


Рис. 5.9. Вы получаете пенсию, пособие, стипендию от государства? [Да - 2386, Нет - 3605]

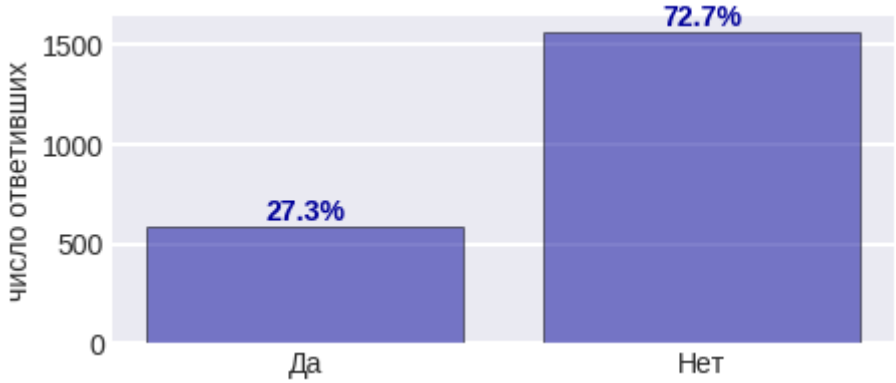


Рис. 5.10. Искали ли Вы работу в течение последних 12 месяцев? [Да - 587, Нет - 1560]

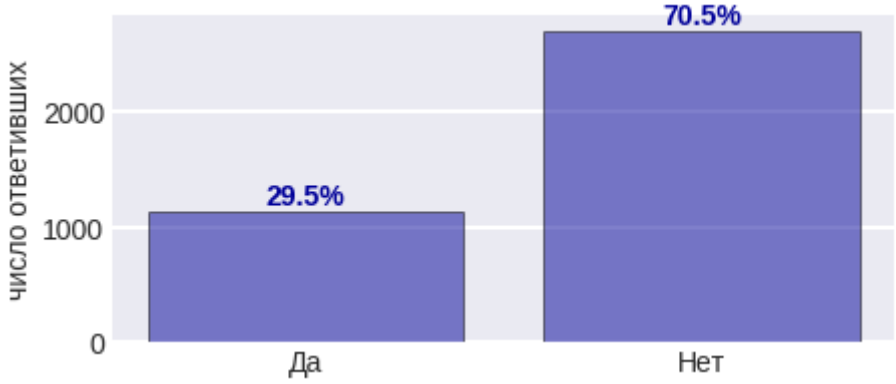


Рис. 5.11. На работе Вы руководите другими людьми? [Да - 1132, Нет - 2702]

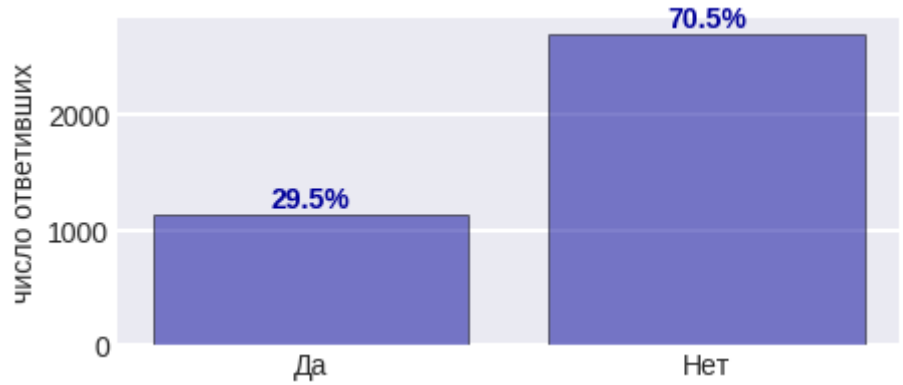


Рис. 5.12. Ваша работа связана с физическим трудом? [Да - 1430, Нет - 1248]

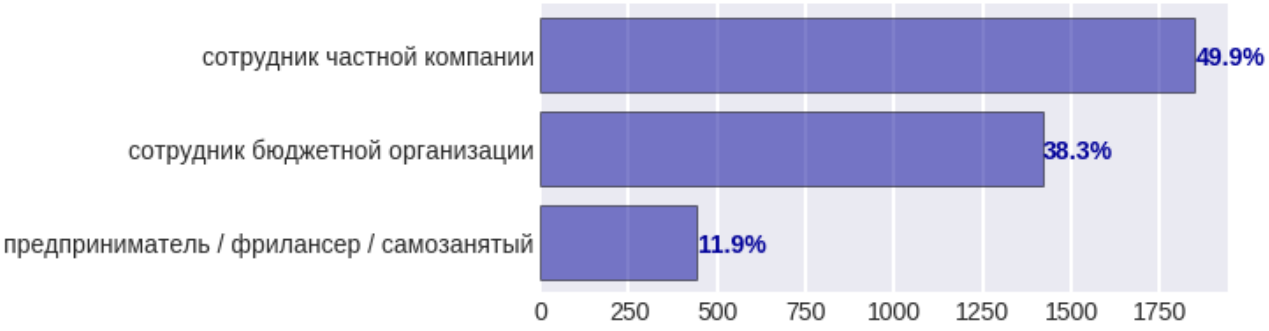


Рис. 5.13. Кем работаете?

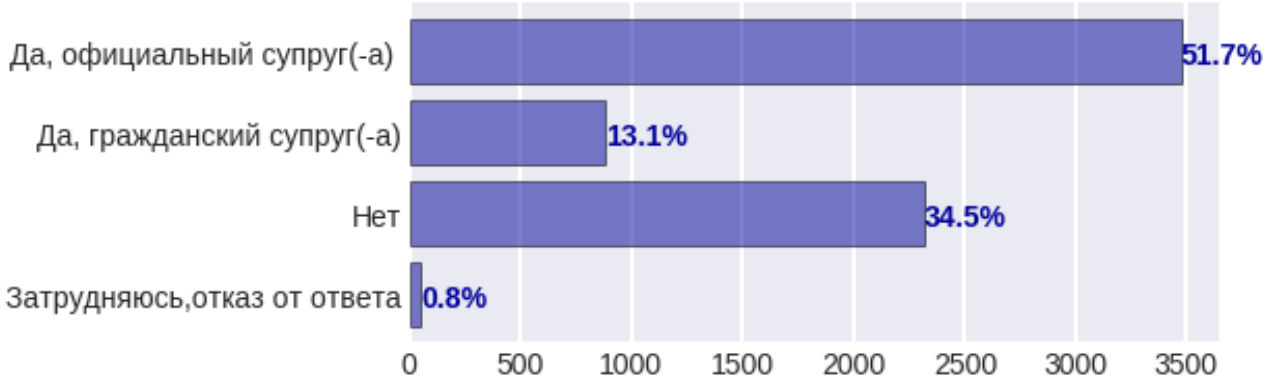


Рис. 5.14. Состоите ли Вы в браке?

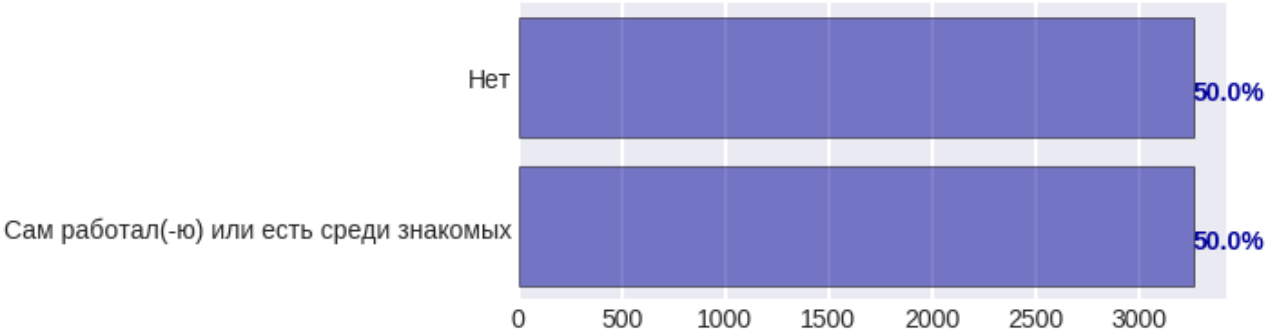


Рис. 5.15. Есть ли среди Ваших знакомых работники правоохранительных органов или Вы сами там работаете или работали?

6. Вид преступления, место и время

Один из вопросов определял, было ли преступление «электронным»: **Q66 «Можно ли сказать, что в Вашем случае преступление произошло через телефон / интернет (например, злоумышленники Вам звонили или писали с просьбой перевести им деньги) или к Вашему случаю это неприменимо?»** Статистика ответов показана в таблице: около 30% электронных преступлений.

	ответов	процентов
Нет, не через телефон или интернет	2075	69.1%
Да, через телефон или интернет	903	30.1%
Затрудняюсь ответить	23	0.8%

Табл.6.1. Ответы на вопрос Q66 «Можно ли сказать, что в Вашем случае преступление произошло через телефон / интернет (например, злоумышленники Вам звонили или писали с просьбой перевести им деньги) или к Вашему случаю это неприменимо?»

Статистика «электронных» преступлений вполне естественна, например на рис. 6.1 показано распределение ответов на вопрос **Q1414** ***«Где Вы в этот момент находились? (преступление через телефон или Интернет)»***, которая вполне соответствует статистике нашего пребывания в различных местах в течение недели.

Рис. 6.1. Распределение ответов о нахождении в момент электронных преступлений.

Интереснее «неэлектронные» преступления, которые подразумевают нахождение «на чужой территории» и/или контакт с преступником.

Распределение ответов на вопрос **Q14** ***«Где произошло преступление?»*** показано на рис. 6.2.

Рис. 6.2. Распределение ответов о месте преступления.

Здесь выбрана довольно подробная шкала, при желании её можно упростить. Например, «квартира/дом» и «дача» – это семейное пространство, «работа/учёба» и «общественные здания» – пространство, которое мы вынуждены посещать, а «улица», «подъезд/двор» – дорога между этими локациями, тогда станет понятно, что примерно треть преступлений застаёт нас в дороге!

Теперь проанализируем время преступления. В анкете был вопрос ***Q16 «В каком примерно месяце это было?»*** . Отметим, что 20.7% опрошенных не помнят точно месяц преступления. Для тех, кто помнит, распределение преступлений по месяцам показано на рис. 6.3. На рис. 6.3 видны «сезонные колебания»: ослабевание преступной деятельности летом (точнее, с мая по июнь – мы потом вернёмся к этому с критической точки зрения). Можно выдвинуть гипотезу, что «электронные» преступления и остальные имеют разные распределения по месяцам, но на рис. 6.4 показано, что это не так (проценты указаны от числа преступлений соответствующего типа).

Если посмотреть на процент «электронных» преступлений (от общего числа преступлений), см. рис. 6.5, то здесь также видны закономерности: с мая по сентябрь их процент (а не только абсолютное число) падает. Отметим, что здесь процент вычислялся не совсем от общего числа, а от числа преступлений, для которых чётко указали электронные они или нет.

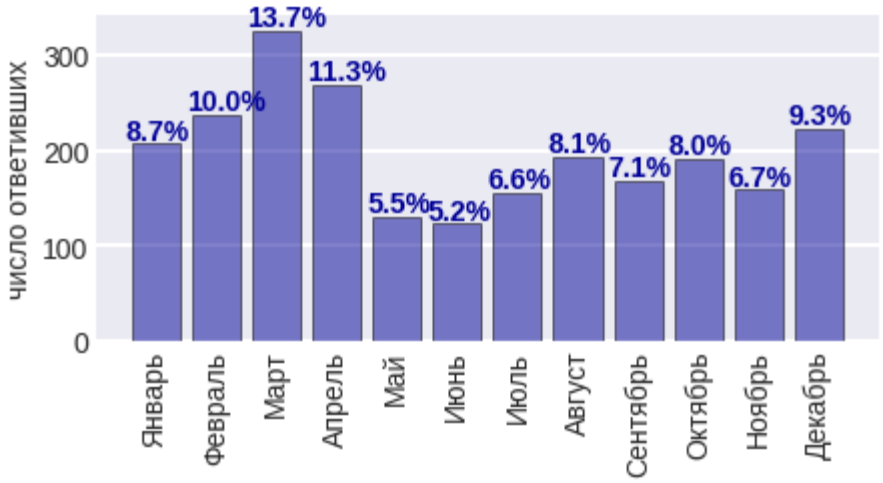


Рис. 6.3. Распределение числа преступлений по месяцам.

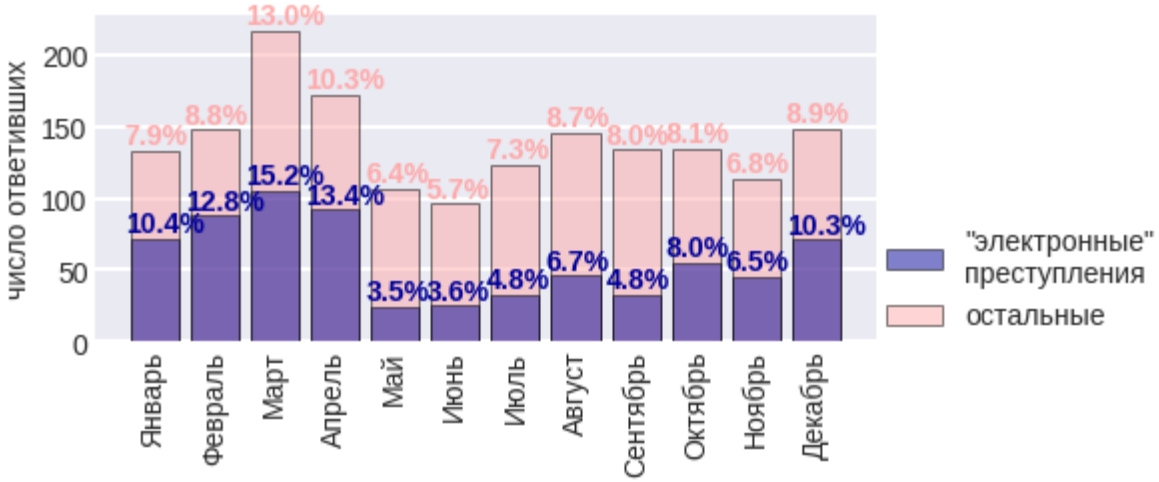


Рис. 6.4. Распределение числа преступлений в зависимости от типа по месяцам.

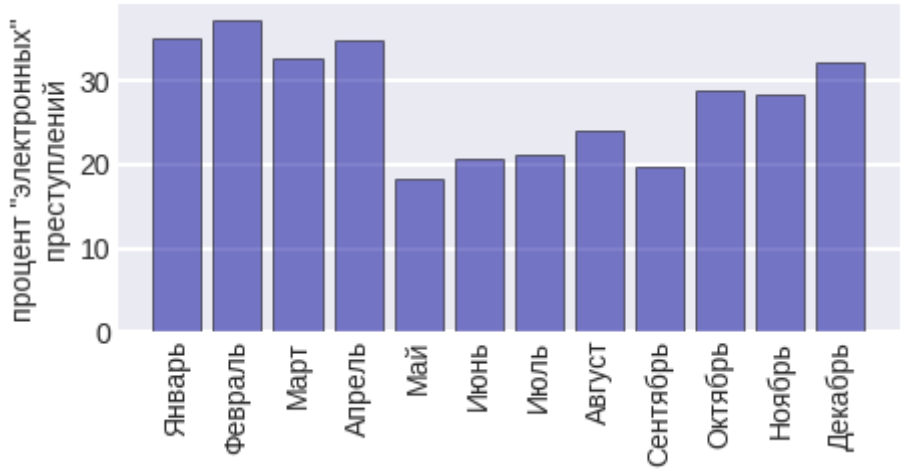


Рис. 6.5. Процент преступлений через мобильный и/или интернет в каждом месяце.

Теперь самое интересное: у тех, кто не помнил месяц преступления спросили про время года, 13.5% из них не помнили и время года, а вот ответы тех, кто назвал время года распределились так, что в каждом сезоне примерно поровну преступлений, см. табл. 6.2.

	ответов	процентов
Весна	136	22.6%
Осень	134	22.3%
Лето	134	22.3%
Зима	117	19.4%
не помню	81	13.5%

Табл. 6.2. Распределение ответов про время года преступления

Но если посмотреть на ответы помнивших точно месяц, то весна – самое насыщенное преступлениями время года, а лето, наоборот, ненасыщенное (и разница довольно очевидна), см. рис. 6.6. Здесь перевод ответа «месяц» во «время года» осуществлялся по принятой календарной схеме, например лето – это июнь, июль и август. Статистики по тем, кто помнит лишь время года не очень много, но это всё равно вызывает недоверие к их показаниям (в смысле абсолютной точности).

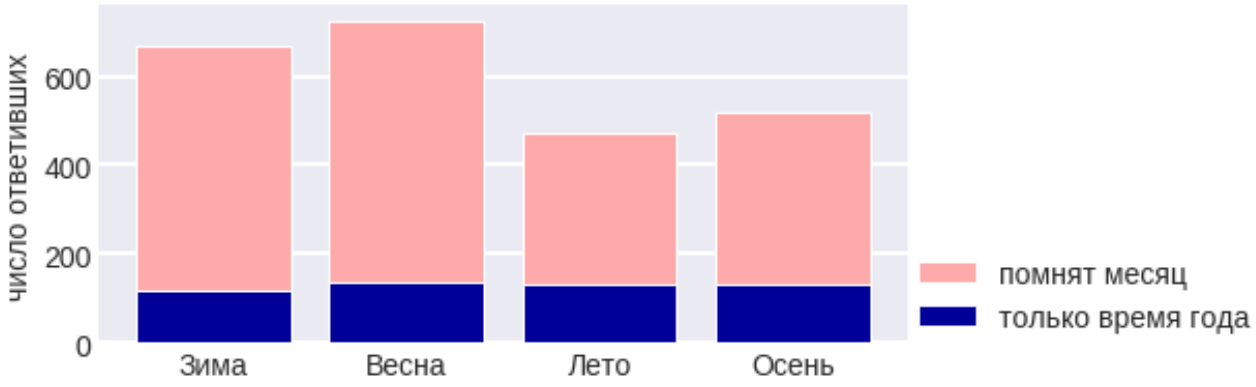


Рис. 6.6. Распределение ответов про время года преступления тех, кто точно помнил месяц и тех, кто точно помнил лишь время года.

Автор подумал, что такая видимая «сезонность» преступлений может быть связана с тем, что опрос проводился где-то в середине-конце апреля, тогда респонденты точно помнили месяцы недавних преступлений (февраль-апрель) и не помнили месяцы давнишних (май-июль). Действительно, в [1] написано, что опрос проводился весной 2018 года: с марта по май. **Очень важно** проводить такие опросы не в фиксированный временной отрезок, а на протяжении всего года, при этом точно указывать дату опроса!

7. Материальный ущерб

Распределение ответов на вопросы про объекты материального ущерба показаны на рис. 7.1. Это была серия вопросов, поэтому показана агрегация их ответов (как часто на каждый вопрос, написанный на оси Y, отвечали «Да»). В целом, вывод очевиден: самый «популярный» материальный ущерб – деньги.

Рис. 7.1. Распределение ответов на вопросы про материальный ущерб.

Рис. 7.2. Распределение ответов на вопрос «Каким вашим имуществом завладели»

Материальный ущерб может заключаться в том, что Вашим имуществом завладели, см. рис. 7.2, или его испортили, см. рис. 7.3. По порче довольно мало статистики, что уберегает нас от выводов, а вот завладевают чаще деньгами.

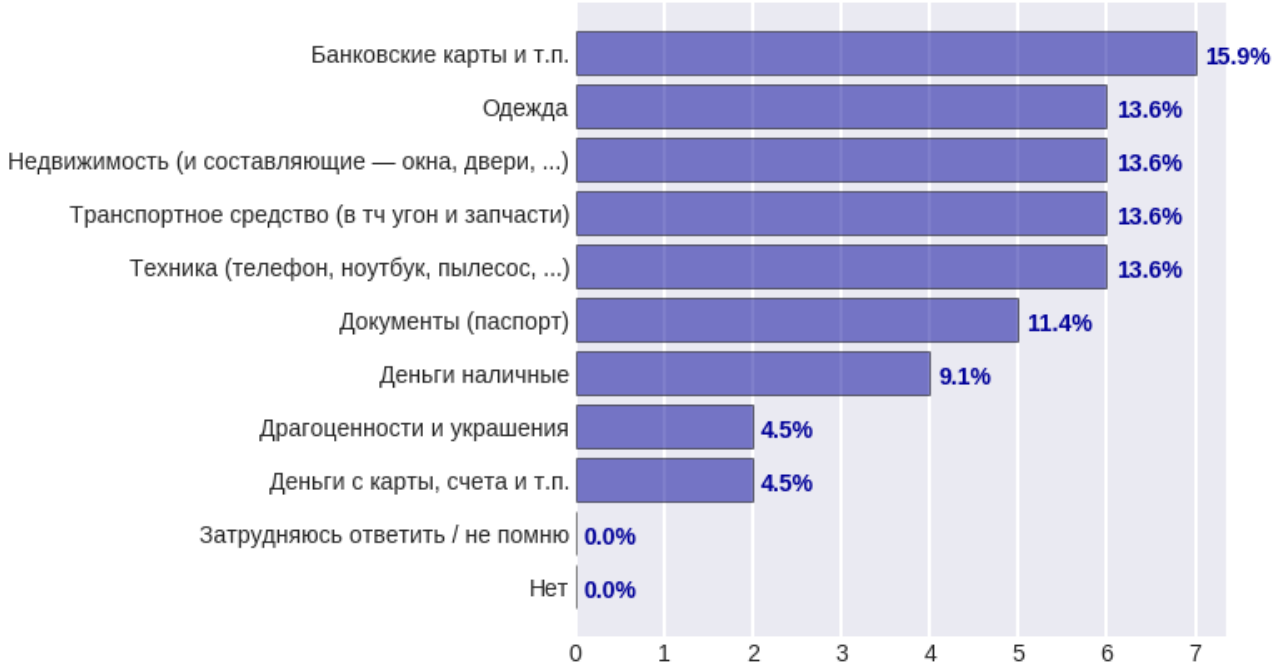


Рис. 7.3. Распределение ответов на вопрос «Какое ваше имущество испортили»

При оценке материального ущерба, как с возрастом, круглые суммы назывались чаще, вот список самых популярных сумм:

сумма	упоминаний
10000	148
5000	139
20000	103
30000	96
15000	88

Максимальная названная сумма ущерба была 6 млрд. руб. Описание преступления было «мошеничество» (без дополнительных подробностей, ID= 14749566). Всего случаев, когда сумма ущерба превышала 0.5 млн. руб – 76, на остальных распределение суммы ущерба показано на рис. 7.4.

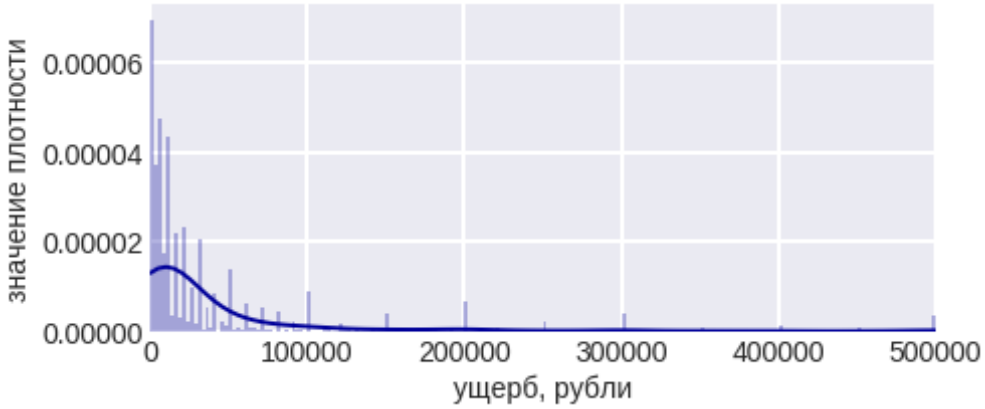


Рис. 7.4. Распределение ущерба респондентов

Отметим, что ущерб фиксировался только, когда был. Например, довольно много электронных преступлений без материального ущерба (по логике он здесь единственно возможный), ниже описания некоторых (часто это СМС-мошенничество, на которое респонденты не велись).

ID=4, 'мошенничество, говорят и звонят я твой сын и дочь, высылать по адресу деньги просят '
ID=8, 'выставыли объявление о продаже квартиры и поступил звонок мне с просьбой они предложили внести аванс на мою карту сбербанка и я продиктовала номер моей карты и они её взломали и получили доступ к моей карте и просили разблокировать карту '
ID=12, ' мошеннические действия'

Также обратим внимание, что медиана «электронных преступлений» – 5600, а остальных – 15000.

8. Исследование электронных преступлений

Мы уже в разделе 6 обсудили «электронные преступления»: через телефон и/или интернет. Для них также был проведён анализ методами машинного обучения (детали мы опускаем), чтобы понять, какие признаки связаны с тем фактом, что преступление электронное. Понятно, что многие вопросы не задавались, тем, кто стал жертвой «электронного» преступления, например

Q30 ***Был ли у Вас риск погибнуть в результате нападения?***
Q26_1, Q26_2, Q26_3 ***вопросы про физический ущерб, телесные повреждения,***
Q25_1, Q25_2, Q25_3 ***вопросы про оружие***
и т.п.

Эти вопросы были исключены из анализа. Из остального есть следующая интересная находка: женщины чаще становятся жертвами «электронных» преступлений, см. табл. 8.1.

	через телефон или интернет	нет	не знаю	процент
Женский	561	1045	12	34.9%
Мужской	342	1030	11	24.9%

Табл. 8.1. Ответы на вопрос, было ли совершено преступление через телефон и/или интернет мужчин и женщин.

Параллельно было найдено много несоответствий в данных. Например, 15 преступлений были заявлены как произошедшие «через телефон или интернет», но при этом в качестве материального ущерба называлась «техника». Давайте рассмотрим их описания. Первая группа эти преступлений такая (с указанием ID, орфография сохранена)

ID=13281665 кража телефона 2017 год конец марта или начала апреля
ID=13313108 осавил телефон в будке на работе его украли
ID=13350647 из дома украли телефон
ID=13532871 Кража телефона у моего ребенка
ID=13894789 украли телефон
ID=14386228 украли телефон
ID=14321635 отдали планшет в ремонт частному мастеру, который планшет не вернул

Ясно, что здесь респонденты путали «кражу через телефон» и «кражу телефона», поэтому отнесение этих преступлений к «электронным» некорректно. Следующая группа описаний:

ID=12710868 июне 2017
ID=13269402 июнь или июль
ID=13329606 кинули
ID=13795476 мошенничество

Здесь из описаний не понятно, что произошло. Заметим, что часто вместо описаний стоит время преступления. Наконец, последняя группа, действительно, относится к электронным преступлениям, в которых материальный ущерб техника, но таких 4 (возможно, 9, но не 15).

ID=13340066 закакзал телефон прислали не то что надо было это в течении трех дней. через три дня после заказа сказали что посылка пришла. принес курьер. в этот момент был на работе. посылку приняла жена. пришел с работы открыл а там не тот телефон. значительно дешев
ID=14399353 украли паспорт и оформили на меня товарный кредит
ID=14737050 Денежная афера,заказала через интернет телефон,ни телефона ни денег
ID=15039496 украли телефон и через сбербанк онлайн сняли деньги

Аналогично, анализируя вопрос о том, кем был преступник, удалось найти преступления, помеченные как электронные, но вряд ли таковыми являющиеся. Отметим, что если в описании стоит «угрожал», то угроза могла быть по телефону (звонком или СМС), но часть преступлений «драка» или «поджог» точно не являются «электронными»:

ID=13308086 продала дачный учасок, а новый владелец угражал

ID=13341640 не оплата коммунальных платежей квартирантами
ID=13373841 осенью в октябре угрожали оскорбление чести и достоинства
ID=13378559 дали цыплят, половина подохли. Сказала, что буду платить только живых, угрожали, что спалят хату
ID=13446010 3 года назад строили дом и нам не выплатили зарплату
ID=13895276 бывший друг, наркоман, угрожал, заставлял молится перед смертью
ID=13910173 драка
ID=13920894 применили физическое насилие и моральное воздействие
ID=14238688 поджог дома дочки
ID=14317258 Человек на судебном заседании под камеру угрожал расправой
ID=14539078 угрозы. сосед пьющий. полиция приезжала но результата нет. февраль 2018

9. Личность преступника

К сожалению, довольно мало людей (373 человека) ответили, кем был преступник (при этом допустимым был ответ «незнакомец»), поэтому статистику, связанную с этим вопросом сложно интерпретировать и обобщать. Укажем лишь несколько интересных находок. В табл. 9.1 показано, как отвечали на вопрос о личности преступника «**Q10** ***Кем злоумышленника Вам приходился?***» мужчины и женщины.

	Женский	Мужской
Друг, подруга	16 (44.4%)	20 (55.6%)
Другой знакомый	47 (40.5%)	69 (59.5%)
Затрудняюсь ответить / не помню	19 (51.4%)	18 (48.6%)
Коллега	22 (34.4%)	42 (65.6%)
Незнакомец	2 (66.7%)	1 (33.3%)
Родственник	26 (72.2%)	10 (27.8%)
Сожитель	13 (86.7%)	2 (13.3%)
Сосед или соседка	35 (63.6%)	20 (36.4%)
Супруг или супруга	11 (100.0%)	0 (0.0%)

Табл.9.1. Как отвечали на вопрос «Кем злоумышленника Вам приходился?» представители разного пола.

Любопытно, что ни один из мужчин не сказал, что преступление совершила жена, тогда как 11 женщин назвали преступниками своих мужей. Ниже перечислены описания этих преступлений, как правило, это избиение супруги, часто в пьяном виде:

ID=13327123 был пьяный , не понравилось что закрыла двери - забежил и избил
ID=13336563 избиение 2017
ID=13562694 ругательство переходило в драку
ID=13791110 Издевались, избивал

ID=13905457 избил супруг
ID=14282324 муж по пьяни избил. 23 февраля 2018
ID=14310517 выламливал дверь.избил беременную
ID=14318206 угрожал бывший муж
ID=14396579 В Новый Год. Мы подрались с мужем. Он меня ударил по глазу. Я вызвала милицию.
ID=15091056 муж кухонный боксёр.минут 15
ID=15309856 муж угрожал убить

Аналогичная ситуация с сожителями: только у двух мужчин был ответ на вопрос о личности преступника «сожитель», причём, судя по описанию преступления, это мог быть сожитель-мужчина.

Кстати, в кодовой таблице, которая была приложена к данным, найдена ошибка: в вопросе Q10 код «7» соответствует ответу «Сосед или соседка», а код 8 – ответу «Сожитель», а не наоборот, это видно из описания преступлений.

Ещё интересное наблюдение: месяц преступления чаще помнят, если преступник супруг / сожитель, а меньше, если сосед/соседка, см. табл. 9.2 (в ней также запомнили время всех преступлений с незнакомцами, но их было всего 3).

Q16	помнит месяц	не помнит
Друг, подруга	32 (88.9%)	4 (11.1%)
Другой знакомый	99 (85.3%)	17 (14.7%)
Затрудняюсь ответить / не помню	32 (86.5%)	5 (13.5%)
Коллега	56 (87.5%)	8 (12.5%)
Незнакомец	3 (100.0%)	0 (0.0%)
Родственник	32 (88.9%)	4 (11.1%)
Сожитель	14 (93.3%)	1 (6.7%)
Сосед или соседка	42 (76.4%)	13 (23.6%)
Супруг или супруга	10 (90.9%)	1 (9.1%)

Табл.9.2. Как отвечали на вопрос «Кем злоумышленника Вам приходился?» те, кто точно помнит и не помнит месяц преступления.

10. Другие вопросы

Ниже просто визуализации распределений ответов на другие вопросы, они не подвергались поверхностному анализу, как предыдущие. Все они были

автоматически получены с помощью написанной автором библиотеки для визуализации результатов опроса.

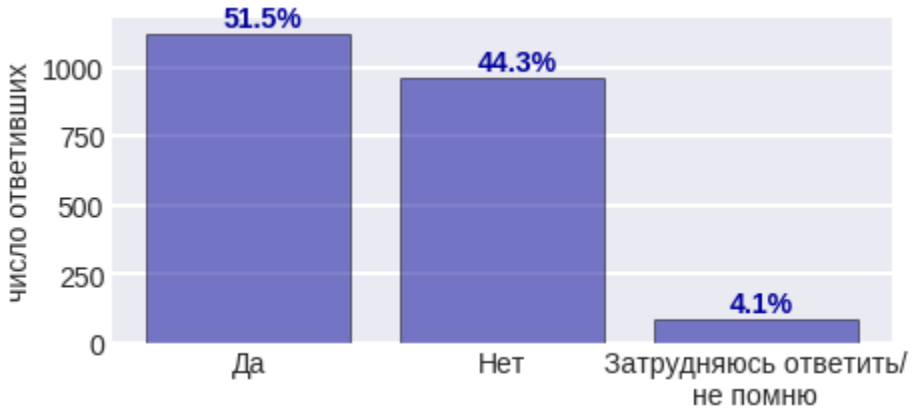


Рис. 10.1. Можно ли сказать, что вас обманули или ввели в заблуждение?
[Да - 1122, Нет - 965, Затрудняюсь ответить/не помню - 90]

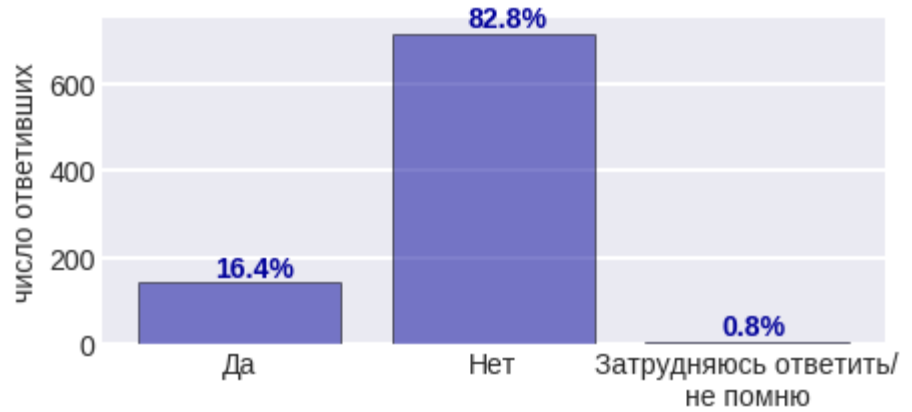


Рис. 10.2. Угрожали ли вам порчей или уничтожением имущества?
[Да - 141, Нет - 714, Затрудняюсь ответить/не помню - 7]

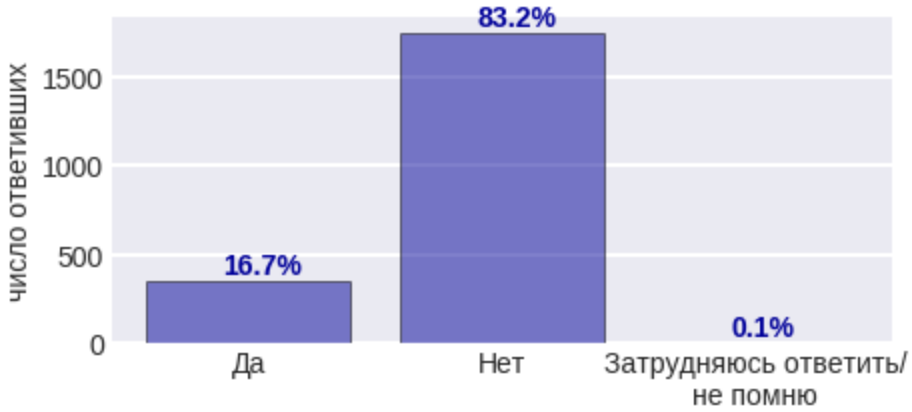


Рис. 10.3. Было ли к Вам применено физическое насилие?
[Да - 350, Нет - 1745, Затрудняюсь ответить/не помню - 3]

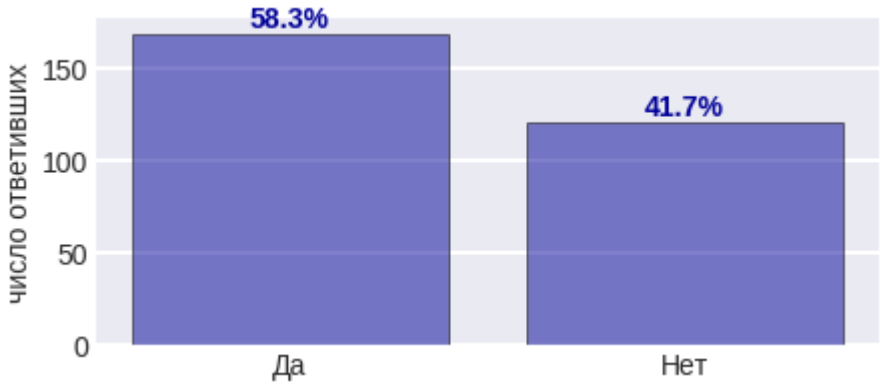


Рис. 10.4. Нуждались ли Вы в медицинской помощи? [Да - 169, Нет - 121]

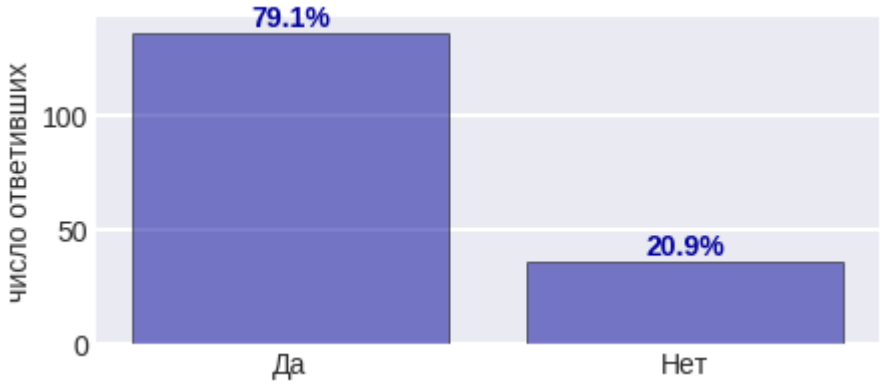


Рис. 10.5. Обращались ли Вы в медицинское учреждение? [Да - 136, Нет - 36]

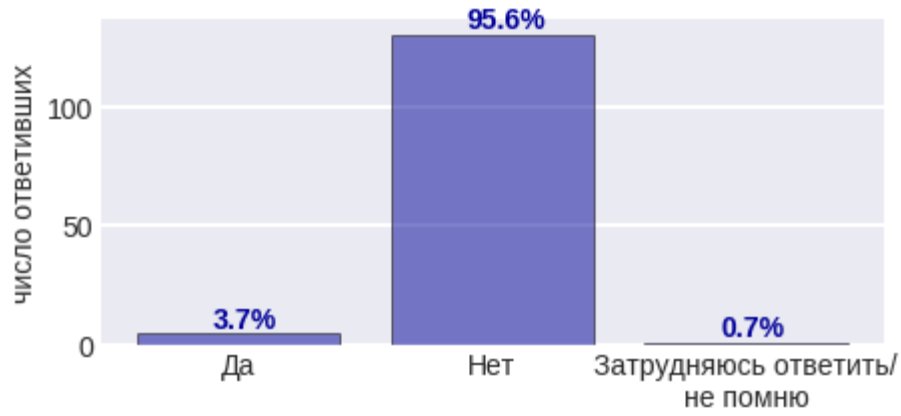


Рис. 10.6. Получили ли вы в результате этого случая инвалидность? [Да - 5, Нет - 130, Затрудняюсь ответить/\nне помню - 1]

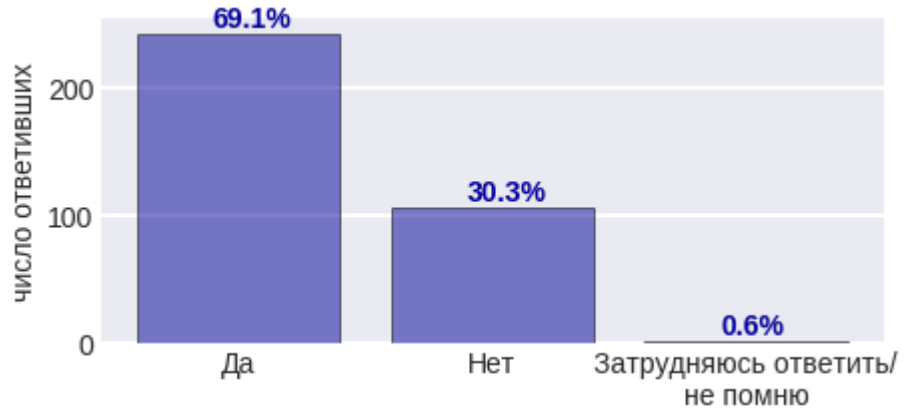


Рис. 10.7. Оказывали ли Вы сопротивление преступнику? [Да - 242, Нет - 106, Затрудняюсь ответить/не помню - 2]

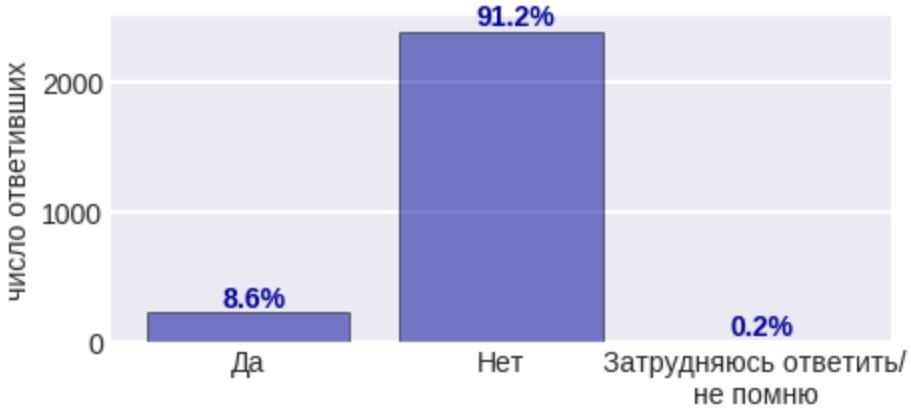


Рис. 10.8. Угрожали ли Вам в этой ситуации физическим насилием? [Да - 226, Нет - 2391, Затрудняюсь ответить/не помню - 4]

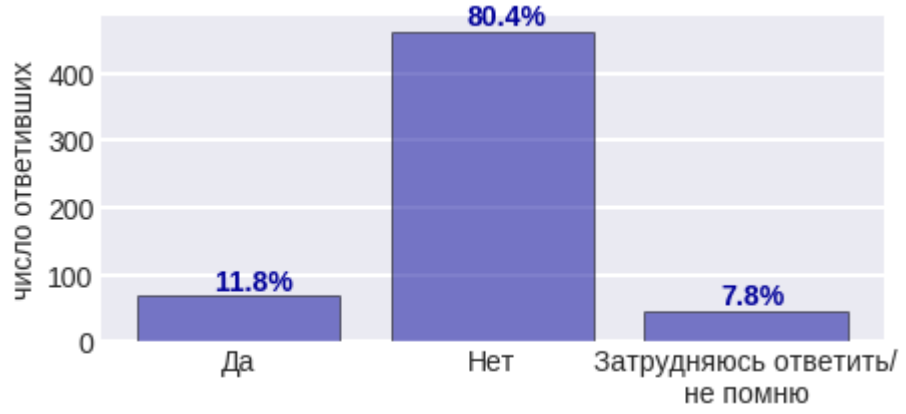


Рис. 10.9. Как вы думаете, могли быть у злоумышленника сексуальные мотивы? [Да - 68, Нет - 463, Затрудняюсь ответить/не помню - 45]

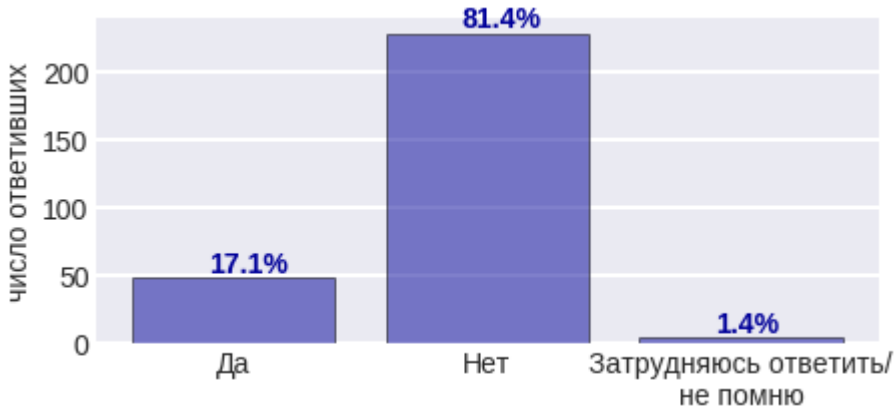


Рис. 10.10. Угрожали ли Вам оружием? [Да - 48, Нет - 228, Затрудняюсь ответить/не помню - 4]

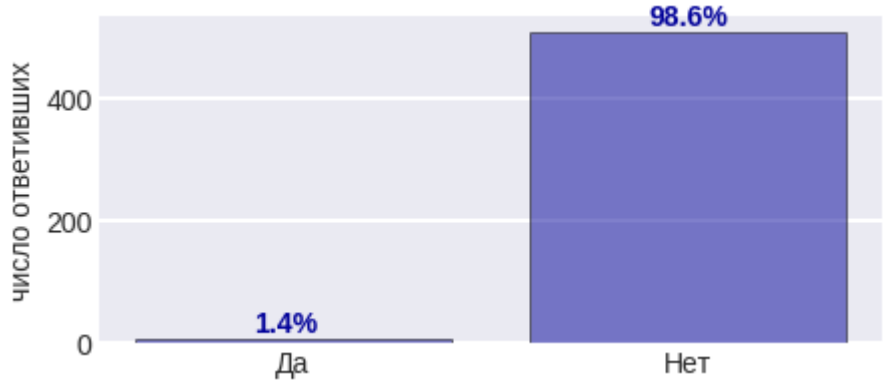


Рис. 10.11. Были ли Вы в той ситуации жертвой ДТП? [Да - 7, Нет - 509]

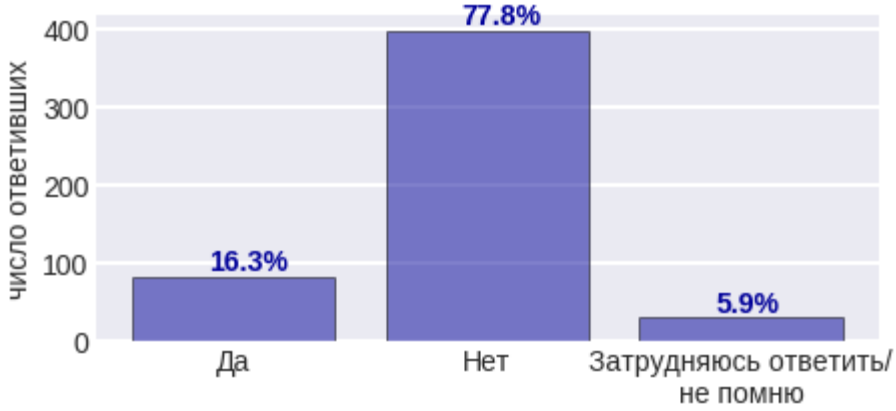


Рис. 10.12. Были ли Вы в той ситуации жертвой чьей-то халатности? [Да - 83, Нет - 397, Затрудняюсь ответить/не помню - 30]

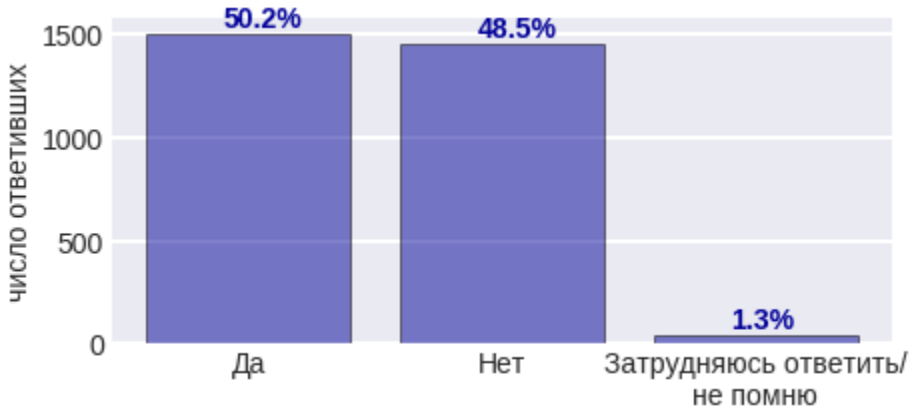


Рис. 10.13. В момент преступления Вы видели злоумышленника или разговаривали с ним? [Да - 1506, Нет - 1455, Затрудняюсь ответить/не помню - 40]

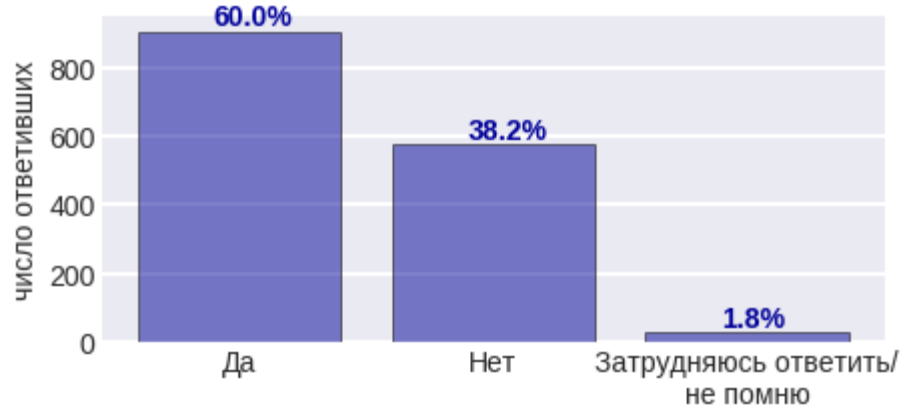


Рис. 10.14. Злоумышленник был один? [Да - 904, Нет - 576, Затрудняюсь ответить/не помню - 27]

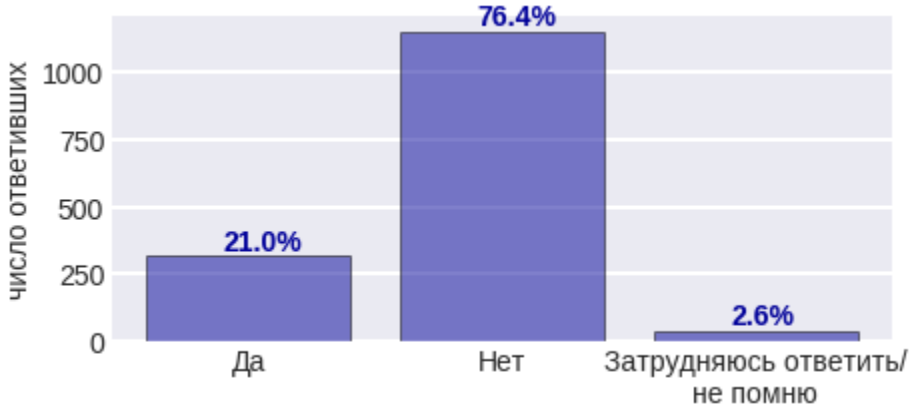


Рис. 10.15. Злоумышленник был мужчина? [Да - 317, Нет - 1150, Затрудняюсь ответить/не помню - 39]

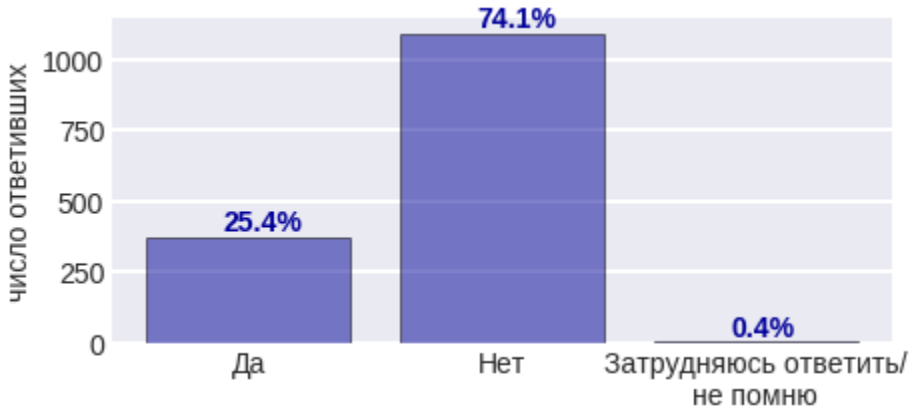


Рис. 10.16. Были ли вы знакомы со злоумышленником? [Да - 374, Нет - 1090, Затрудняюсь ответить/не помню - 6]

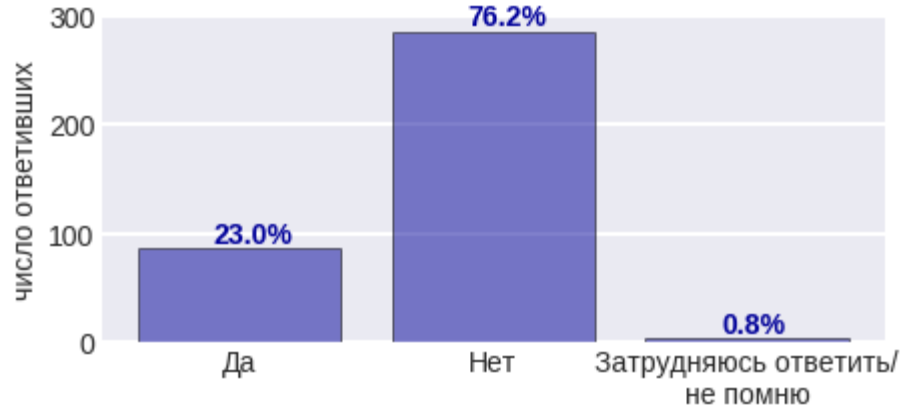


Рис. 10.17. Можно ли сказать, что Вы находились в зависимом положении от преступника (он был начальником, основным кормильцем или старшим родственником)? [Да - 86, Нет - 285, Затрудняюсь ответить/не помню - 3]

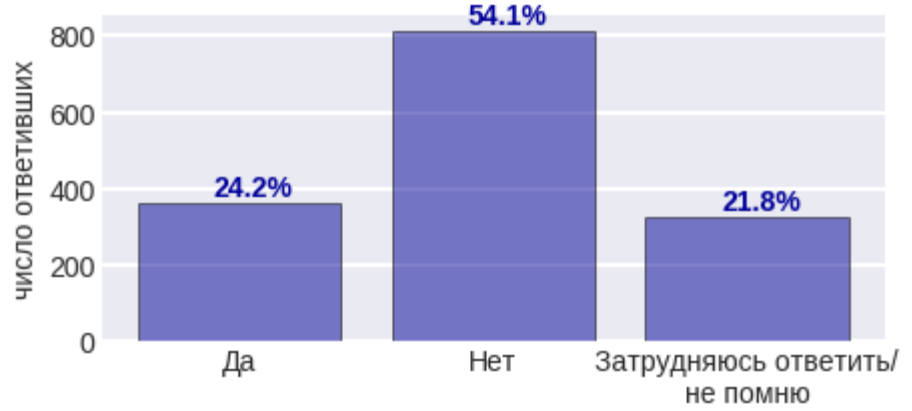


Рис. 10.18. Был ли преступник или кто-нибудь из преступников в состоянии алкогольного или наркотического опьянения? [Да - 364, Нет - 814, Затрудняюсь ответить/\nне помню - 328]

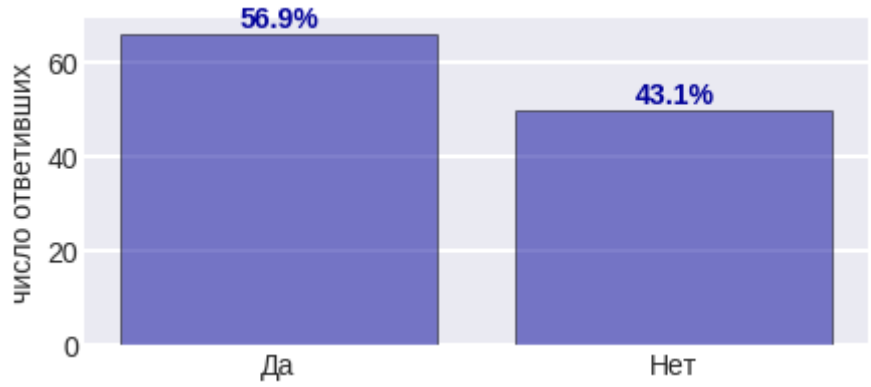


Рис. 10.19. Преступление, которое мы обсуждали, было единственным или повторялось (случилось не меньше двух раз)? [Да - 66, Нет - 50]

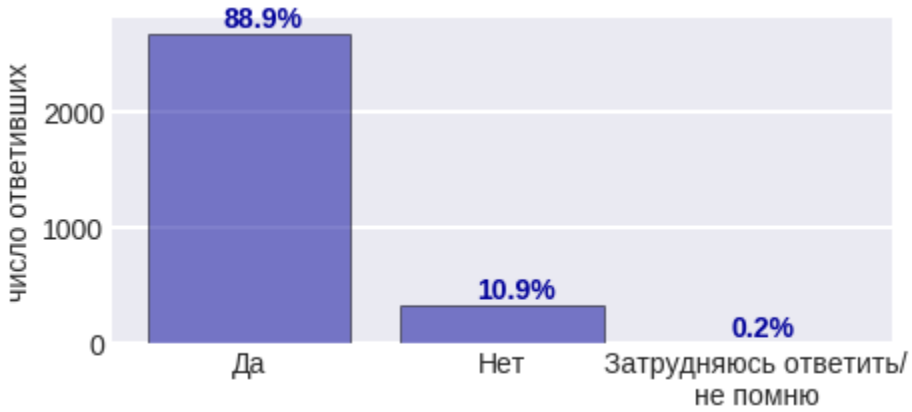


Рис. 10.20. Рассказывали ли Вы кому-нибудь из близких о случившемся? [Да - 2668, Нет - 328, Затрудняюсь ответить/не помню - 5]

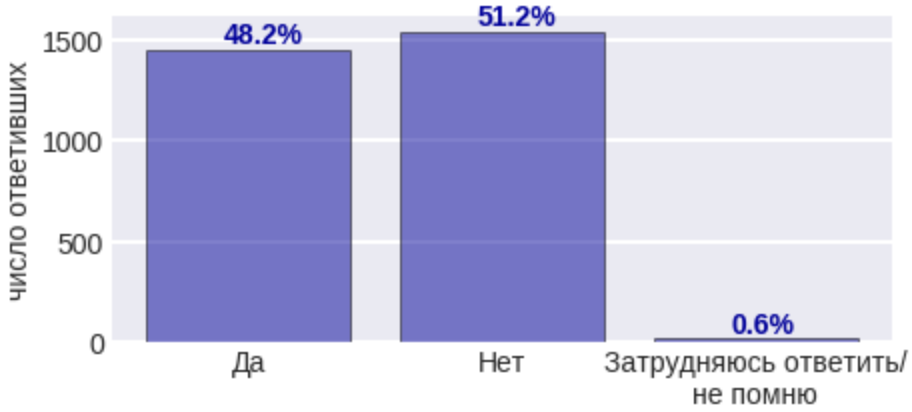


Рис. 10.21. Узнали ли о случившемся правоохранительные органы? [Да - 1447, Нет - 1537, Затрудняюсь ответить/не помню - 17]

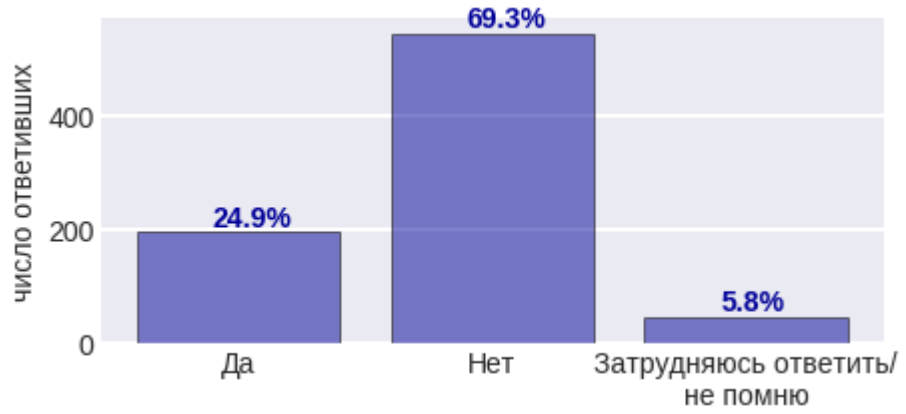


Рис. 10.22. Примирились ли Вы с виновными? [Да - 196, Нет - 545, Затрудняюсь ответить/не помню - 46]

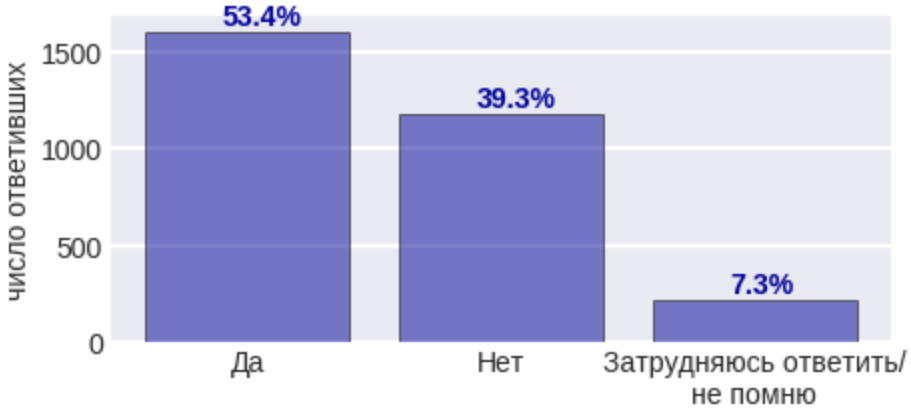


Рис. 10.23. Оказавшись снова в аналогичной ситуации, Вы, скорее всего, обратились бы в полицию? [Да - 1602, Нет - 1179, Затрудняюсь ответить/не помню - 220]

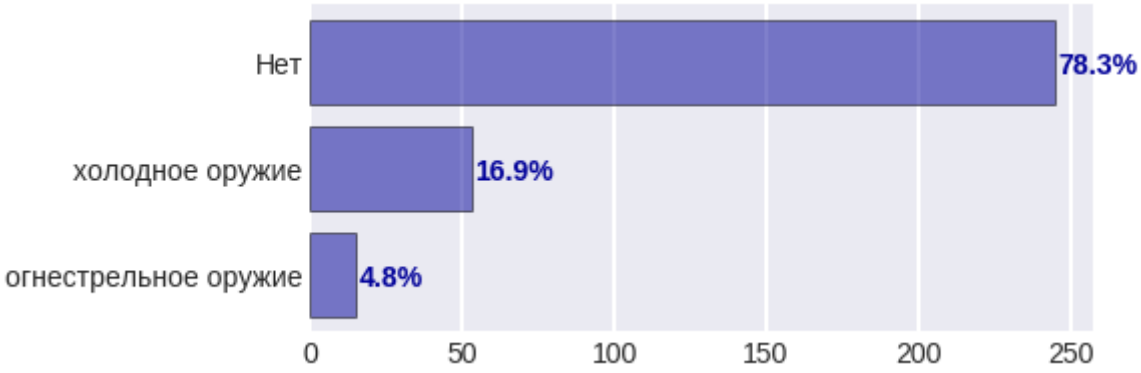


Рис. 10.24. Использовалось ли что-либо в качестве оружия?

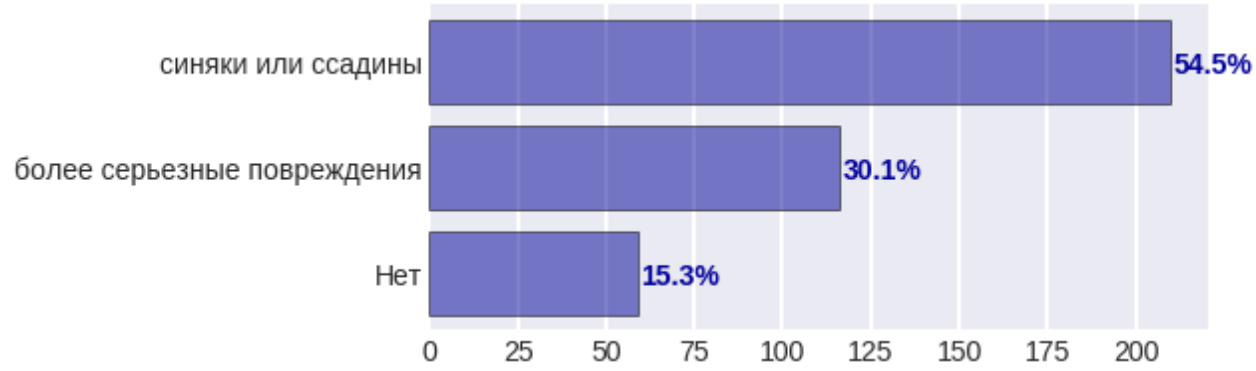


Рис. 10.25. Был ли Вам причинен какой-либо физический ущерб, телесные повреждения?

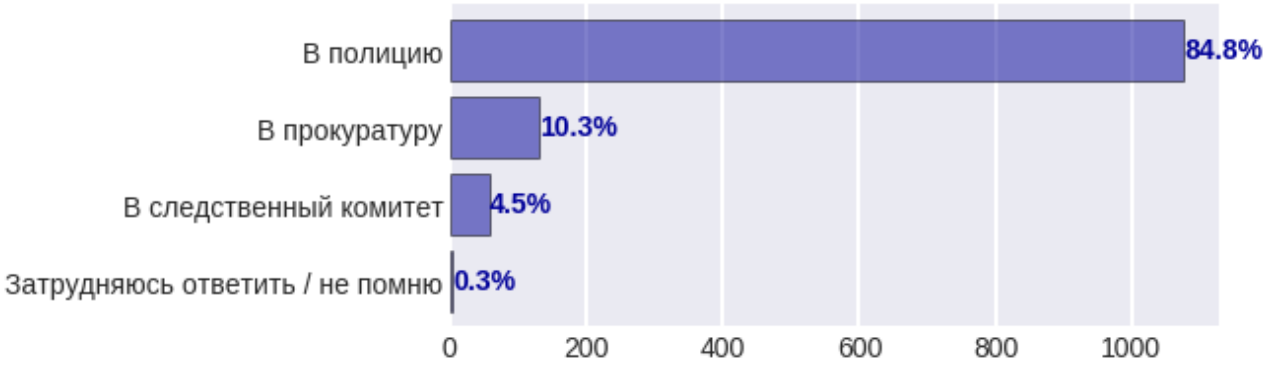


Рис. 10.26. Куда именно Вы обратились?

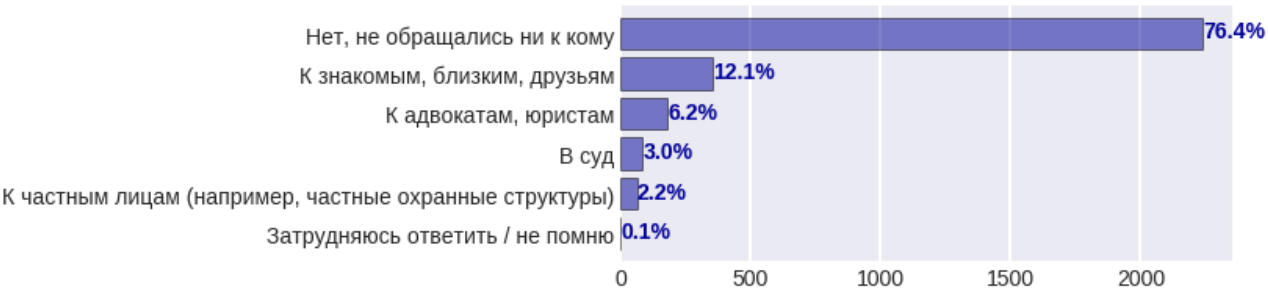


Рис. 10.27. Обращались ли Вы за помощью по поводу случившегося еще к кому-нибудь, кроме правоохранительных органов?

Рис. 10.28. Преступник был должностным лицом или при исполнении служебных обязанностей?

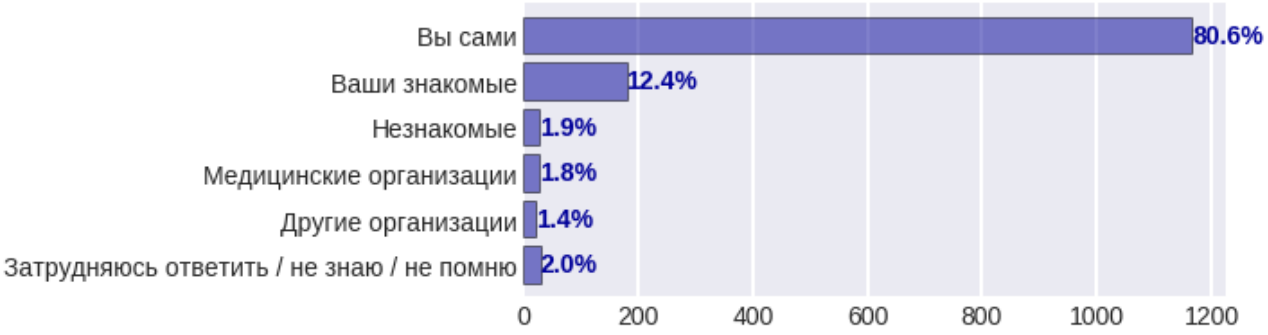


Рис. 10.29. Кто обратился в правоохранительные органы?

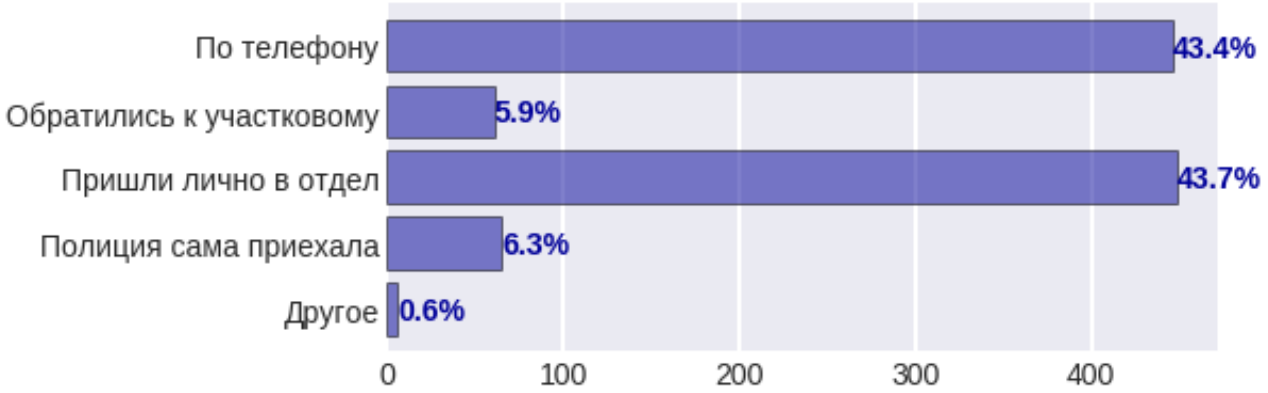


Рис. 10.30. Как именно Вы обратились в полицию?

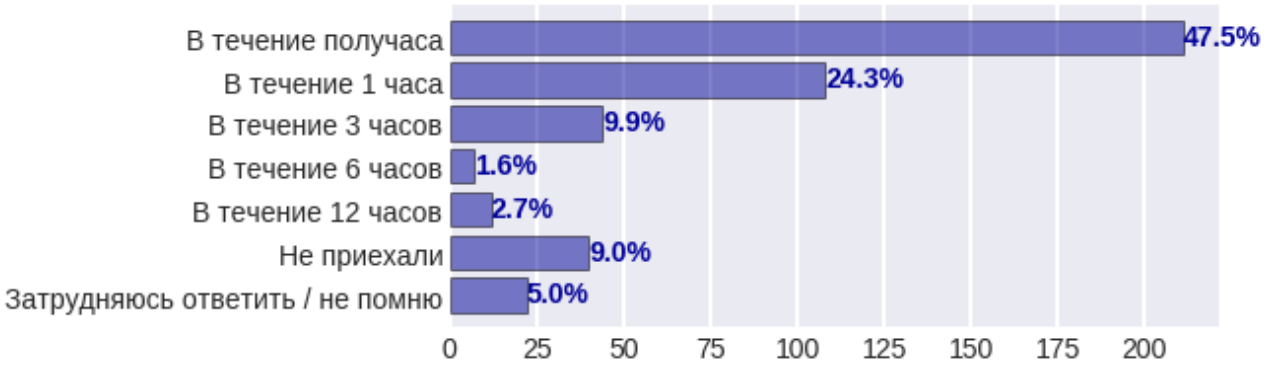


Рис. 10.31. Как быстро полиция приехала после Вашего звонка?

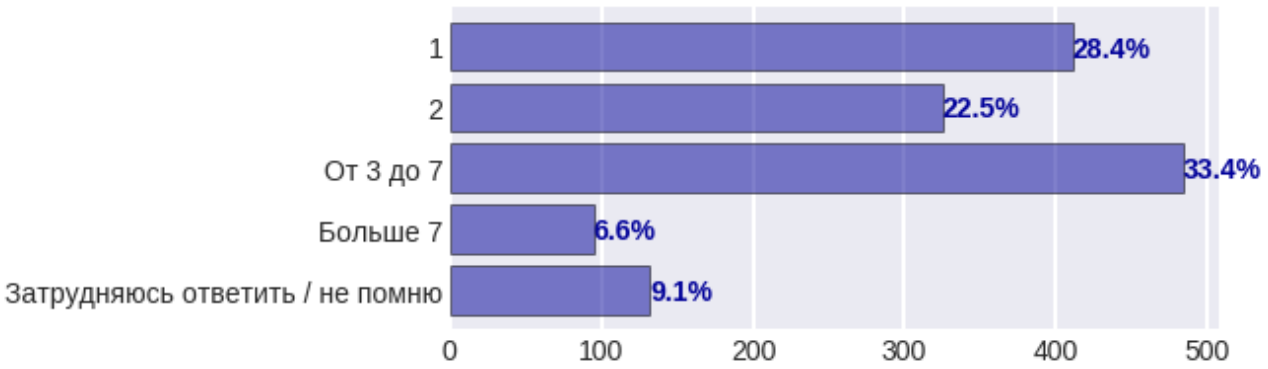


Рис. 10.32. Сколько раз Вы встречались с сотрудниками правоохранительных органов?

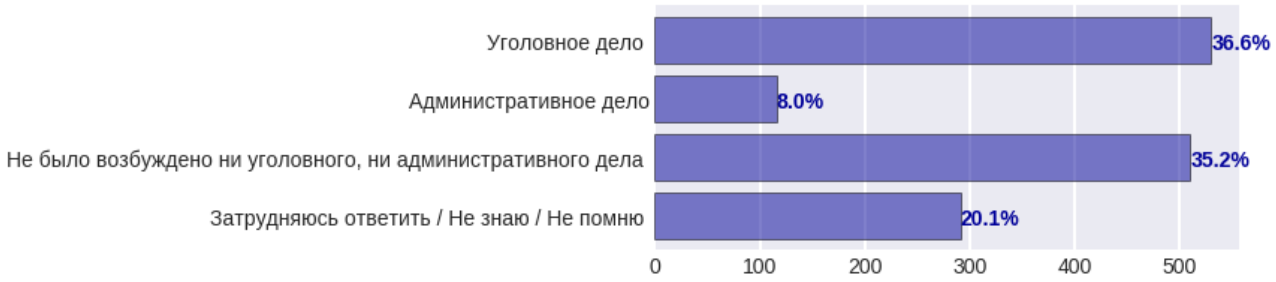


Рис. 10.33. В результате, они возбудили уголовное или административное дело?

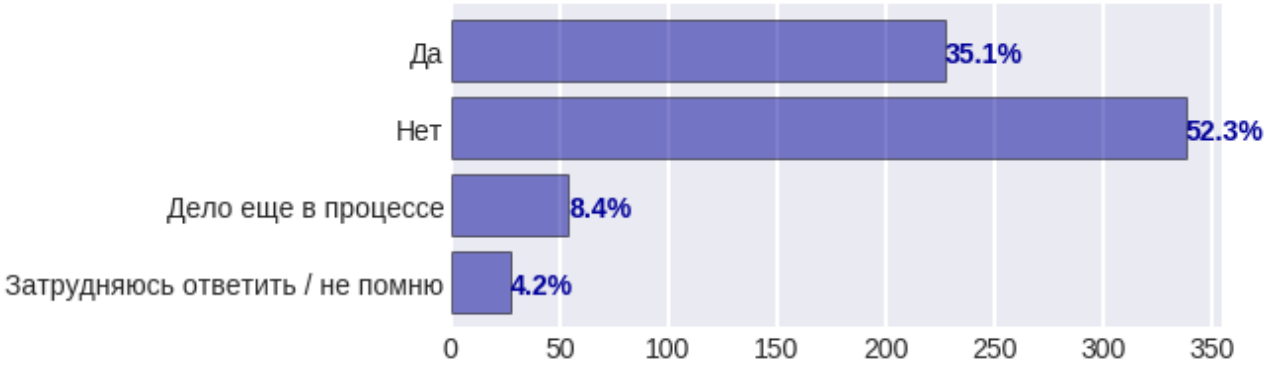


Рис. 10.34. Дошло ли дело до суда?

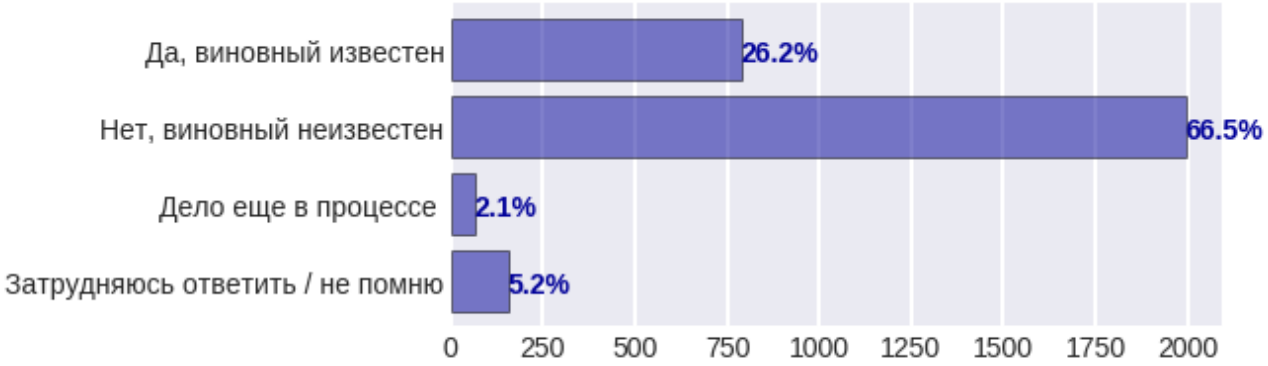


Рис. 10.35. Был ли установлен виновный или кто-либо из виновных, если их было несколько?

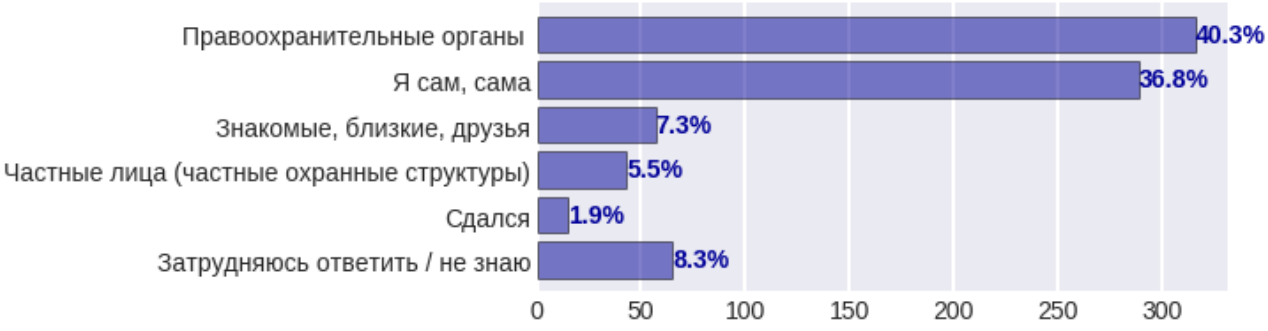


Рис. 10.36. Кем был найден виновный или кто-либо из виновных, если их было несколько?

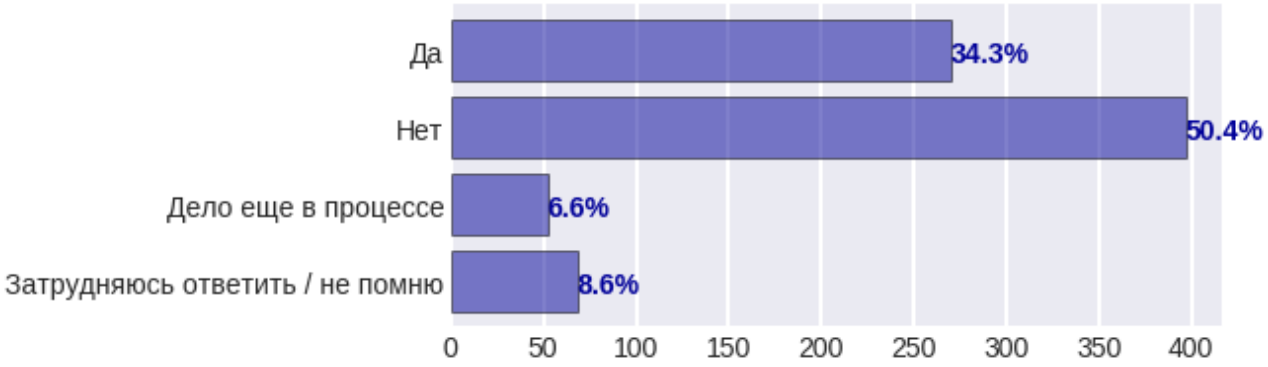


Рис. 10.37. Получил ли наказание виновный или кто-либо из виновных, если их было несколько?

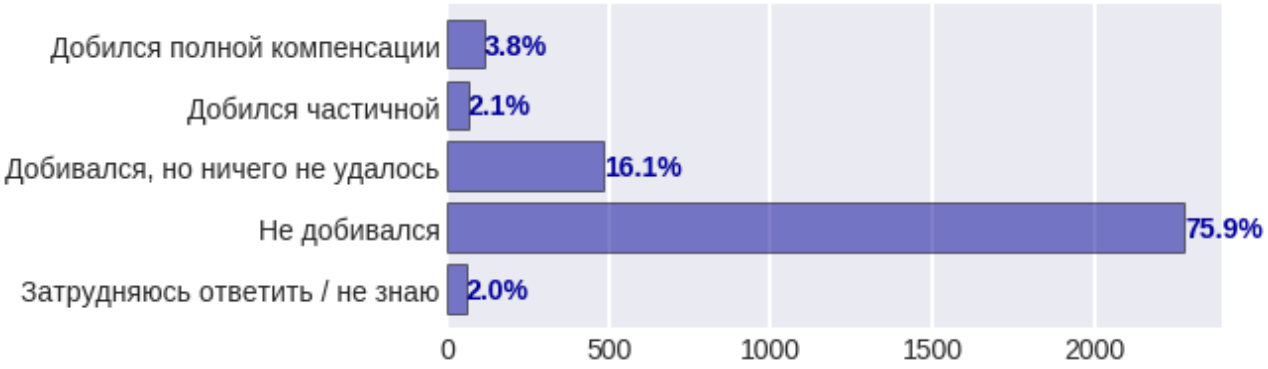


Рис. 10.38. Добивались ли Вы материальной компенсации? Если да, то удалось ли ее добиться?

Заключение

В результате анализа данных ревиктимизационного опроса написана библиотека для визуализации ответов на вопросы, в отчёте представлены визуализации почти всех ответов. Библиотека выложена в общий доступ.

Была построена модель алгоритмов машинного обучения для определения, являлся ли респондент жертвой преступления по данным его анкеты. С помощью этого алгоритма выявлены ключевые признаки в анкете: возраст, судимость, одиночное проживание и т.п. (показано, что проценты жертв различаются в группах респондентов с разными значениями этих признаков).

Получены интересные гипотезы, например, что люди, которые уклоняются от определённых ответов, с большей вероятностью не были жертвами преступлений (требуется проверка на большем массиве данных); люди, которые точно не помнят месяц преступления, скорее всего, ошибутся и со временем года (также нужна проверка); есть сезонные колебания в числе

преступлений и в проценте «электронных» преступлений (нужна проверка с помощью опроса, который проводится в течение всего года).

Выявлены некоторые недостатки опроса. Например, его проведение в течение небольшого промежутка времени в конце весны. В результате, респонденты лучше помнили весенние и зимние преступления (возможно, что у летних и осенних преступлений также больше неточностей в описаниях).

Получены интересные для автора наблюдения, например, что чем выше уровень образования, тем больший процент респондентов с таким образованием был жертвой преступлений; а женщины чаще становятся жертвами «электронных» преступлений.

Также обнаружены многочисленные неточности в данных. Часть из них связана с тем, что респонденты ошибочно понимали вопросы, например путали «преступление через телефон» и «преступление с телефоном». Часть с ошибками внесения информации и неправильными кодовыми таблицами.

По результатам технического отчёта составлен тест (ссылку см. в начале отчёта).

В разделе 10 перечислены визуализации ответов, большая часть которых не подвергалась анализу. В основном, они соответствую вопросам о действиях респондентов во время / после преступлений. Это потенциально хорошая область для дальнейших исследований.

Благодарности

Спасибо Институту проблем правоприменения при Европейском университете в Санкт-Петербурге за подготовку данных и интересный конкурс по их визуализации.

Литература

[1] Веркеев А. М., Волков В. В., Дмитриева А. В., Кнорре А. В., Кудрявцев В. Е., Кузнецова Д. А., Кучаков Р. К., Титаев К. Д., Ходжаева Е. А. **Как изучать жертв преступлений?** // Мониторинг общественного мнения: Экономические и социальные перемены. 2019. No 2. С. 4—31. Verkeev A. M., Volkov V. V., Dzmitryieva A. V., Knorre A. V., Kudryavtsev V. E., Kuznetsova D. A., Kuchakov R. K., Titaev K. D., Khodzhaeva E. A. (2019) How to study victims of a crime? // Monitoring of Public Opinion: Economic and Social Changes. No. 2. P. 4—31. doi: 10.14515/monitoring.2019.2.01