Анализ и прогнозирование вероятных мест преступления в городе Los Angeles

Роман Войтенко

1-Технический университет Молдовы

e-mail: [roman.voitenco@iis.utm.md](mailto:roman.voitenco@iis.utm.md)

**Абстракт**

В данной статье представлен анализ данных преступлениях в городе Los Angeles на основе обширного набора данных, с периода начиная с 2020 по 2023 год. Были произведены различные манипуляции с выбранным датасетом, включая графические представления, выбор наиболее важных переменных, также построение моделей машинного обучения для предсказания предположительного места преступления.

**Введение**

Процесс обработки данных является основополагающей для принятия сложных и обдуманных решений, ведь чтобы сделать хоть какой-то выбор следует проанализировать какие-либо данных, в зависимости от ситуации и необходимой задачи.

В данной работе будет описаны взаимодействия и анализ датасета выбранного на данном сайте: [Data.gov Home - Data.gov](https://data.gov/), официальный сайт Американского правительства с датасетами на различные темы и категории. Сам датасет представляет собой данные о преступлениях в голоде Los Angeles начиная с 2020-2023 год, ссылка на датасет: [Crime Data from 2020 to Present - Catalog](https://catalog.data.gov/dataset/crime-data-from-2020-to-present). Данный датасет был выбран в связи с актуальностью темы, а также относительной симпатией к городу Los Angeles, не смотря на то что данные в собраны начиная с 2020 по 2023 год, данных не так уж и мало, за данный период произошло около 800 тысяч преступлений.

Изучение датасета о преступлениях в городе Los Angeles с 2020 по 2023 год отражает не только важность анализа криминальной статистики, но и позволяет погрузиться в динамику событий в одном из крупнейших городов США. Свыше 800 тысяч документированных инцидентов предоставляют обширную базу для изучения тенденций, паттернов и возможных факторов, влияющих на преступность в этом регионе. Анализ этих данных может помочь не только в понимании преступной ситуации, но и в формировании стратегий превентивных мер и улучшения безопасности города.

Более того, данное исследование может стать основой для будущих аналитических работ, направленных на снижение преступности и разработку политик, ориентированных на улучшение качества жизни горожан. Путем анализа данных о преступлениях в Лос-Анджелесе, можно создать основу для принятия информированных решений в области правоохранительной деятельности и управления городской инфраструктурой.

Таким образом, данное исследование имеет не только академическую, но и практическую значимость, предоставляя информацию, которая может быть использована для формирования политики безопасности и улучшения общественной среды в Лос-Анджелесе.

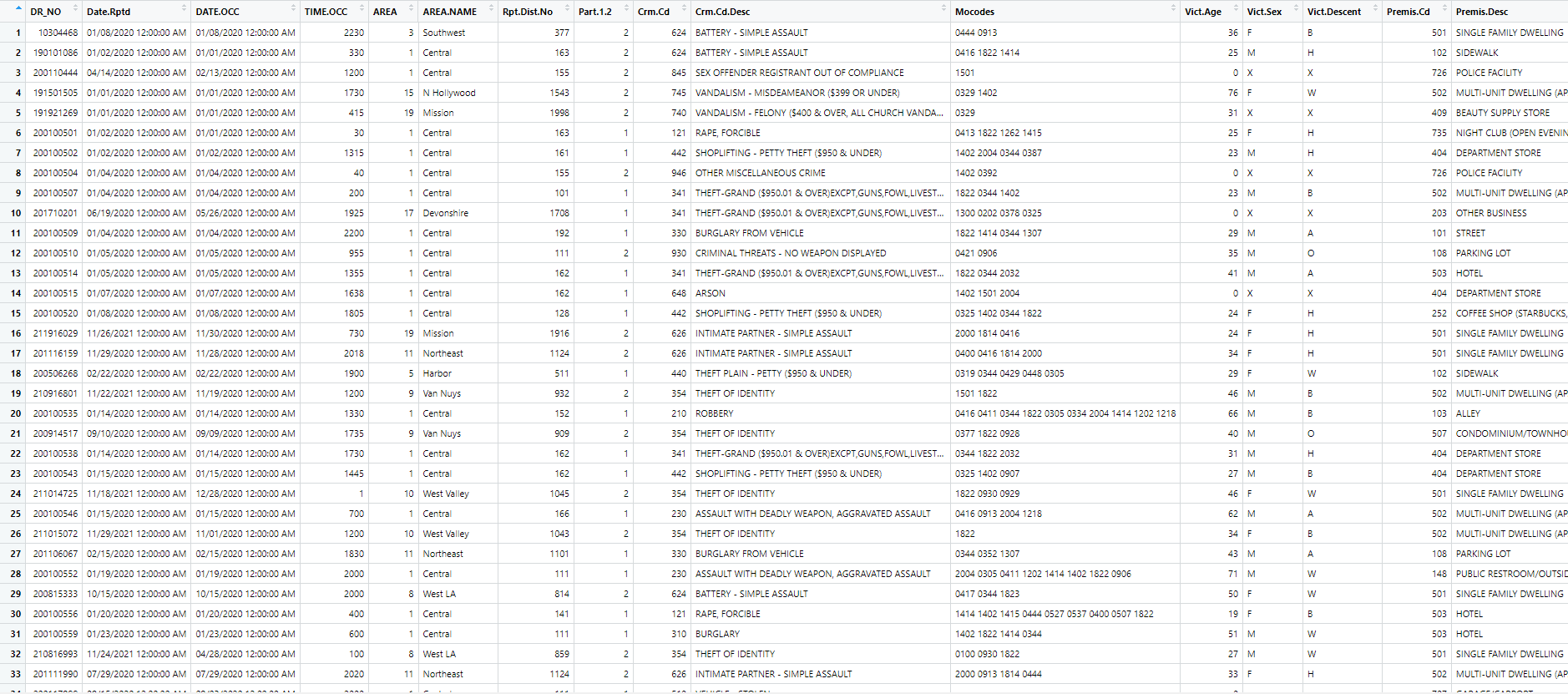
**Материалы и методы**

Для данной работы и использовались следующие материалы и методы: Среда программирования на языке программирования R, то есть сам язык и R studio, благодаря данному инструменту была возможность выводить графики и визуализировать различные ситуации и данные из имеющегося набора данных, также использовались материалы по обучению на полученные на курсе Analiza Datelor, необходимые методы которые проводились на тестовых данных полученные из материалов курса, потом были перенесены на выбранный датасет. Для обучения языка программирования R использовалась официальная документация.

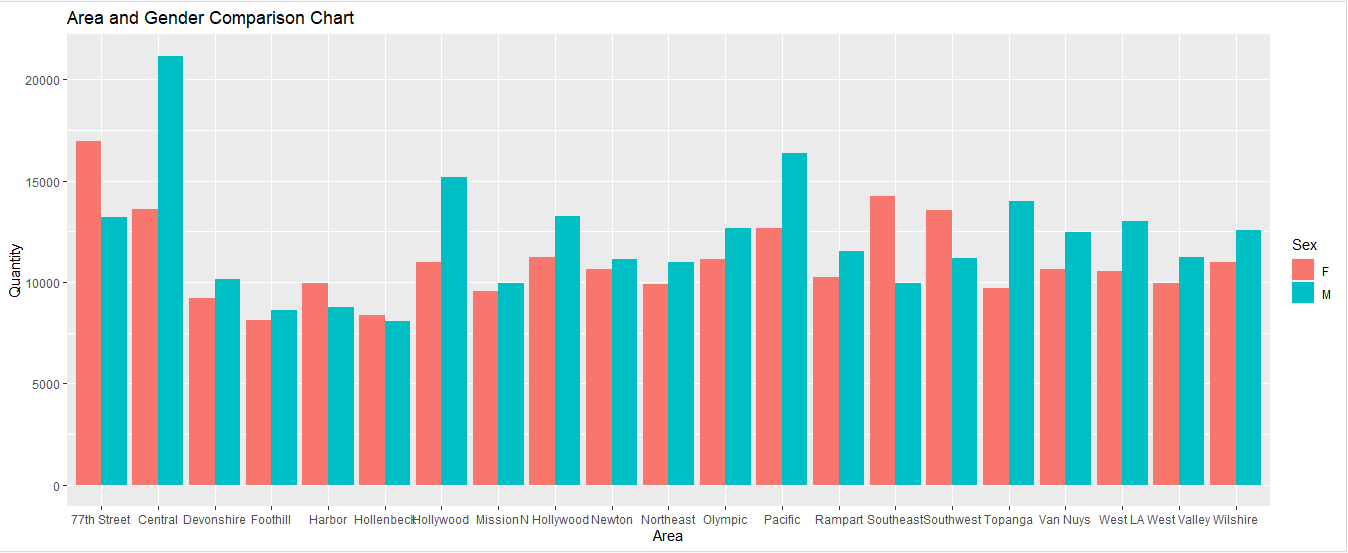
Методы, полученные на курсе по анализу данных, послужили основой для проведения структурированного и основательного исследования данного датасета. Этот опыт обучения и практики на тестовых данных стали фундаментом для применения тех же подходов и техник анализа на реальном датасете о преступлениях. Последовательность использования полученных знаний и методов анализа данных на реальной информации обогатила понимание того, как применять эти инструменты на практике и расширила спектр возможных выводов из исследования.

Для построения модели я использовал линейную регрессию, построение логической модели в планах не было, так как особо не было представления с какими переменными это можно сделать.

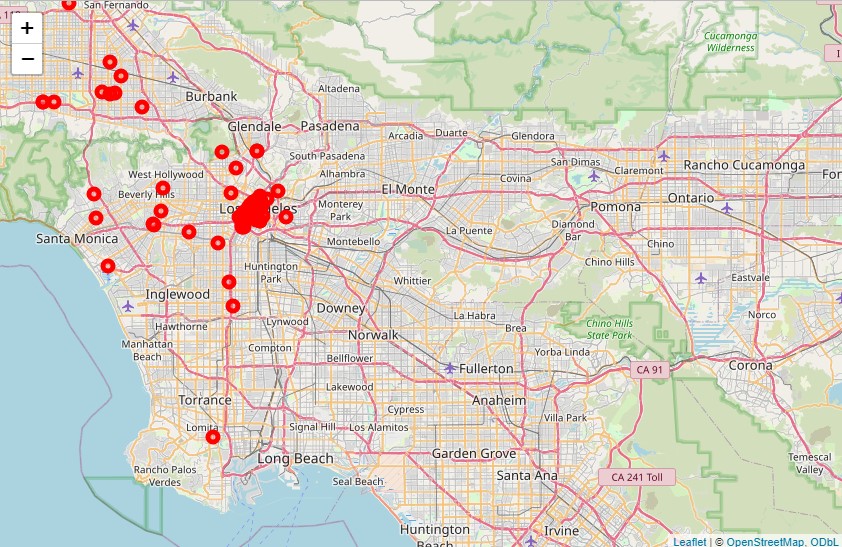
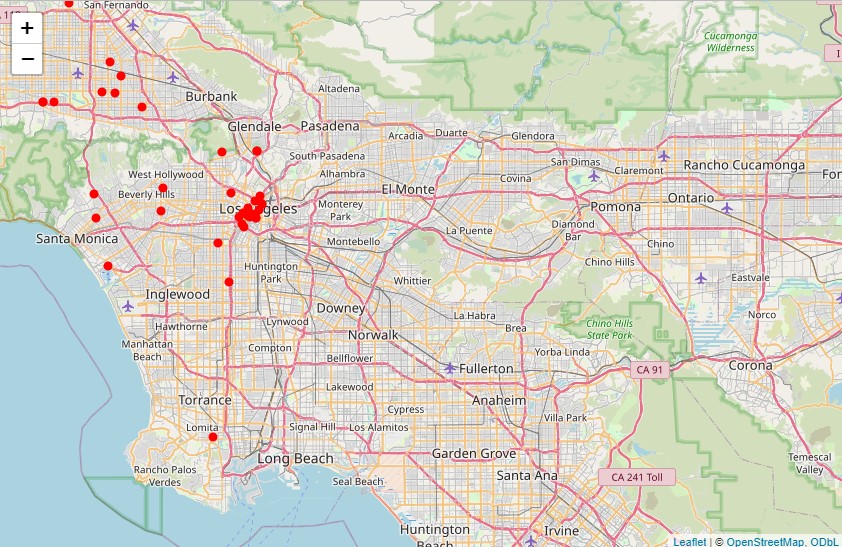
Выбранный датасет содержит в себе 28 переменных, для создания основной модели происходил подбор наиболее подходящих из них. Все переменные из датасета: DR\_NO,Date Rptd,DATE OCC,TIME OCC,AREA,AREA NAME,Rpt Dist No,Part 1-2,Crm Cd,Crm Cd Desc,Mocodes,Vict Age,Vict Sex,Vict Descent,Premis Cd,Premis Desc,Weapon Used Cd,Weapon Desc,Status,Status Desc,Crm Cd 1,Crm Cd 2,Crm Cd 3,Crm Cd 4,LOCATION,Cross Street,LAT,LON



**Рисунок 1**



**Рисунок 2**



**Рисунок 3 Рисунок 4**

Конечно, в процессе обработки данных применялись стандартные методы преобразования строковых данных в числовые, а также удаления пустых значений. Это позволило убедиться, что данные соответствуют нужным критериям для дальнейшего анализа или использования в моделях.

**Результаты**

В ходе изучения и анализа датасета о преступлениях в городе Los Angeles, было получено различное количество данных в зависимости от взаимодействия с датасетом и его переменных. Во-первых, построение модели оказалось не самым простым действием, которое можно было реализовать, были подобраны различные переменные и в зависимости от них точность модели по предсказанию вероятного места преступления варьировалась от 0.001 до 0.06, как видно, реализовать хорошую и точную модель не получилось.

Самая точная модель получилась, используя переменные AREA и Vict.Age, во-первых, птмш это одни из немногих переменных которые было возможно использовать построения регрессивной модели, выше на рисунках 3 и 4, показаны результаты работы модели и как видно, они не особо сильно различаются между собой

Однако, важно отметить, что в процессе анализа мы столкнулись с ограничениями и вызовами, которые могли повлиять на точность модели. Возможно, для дальнейшего улучшения модели потребуется дополнительный анализ данных, использование более сложных моделей или внесение других переменных для учета различных аспектов преступлений в городе.

**Обсуждения**

В ходе выполнения анализа, выбранного датасета, гипотеза о зависимости возраста жертвы от места преступления, доказана не была, так как так и не получилось реализовать модель достаточной точности, чтобы можно было ее использовать в реальной жизнь, хотя бы и в режиме тестирования

Первая модель была построена на широте, долготе и возрасте жертвы, как оказалось для обучения модели данных очень немного и точность была менее одной десятой.

Во второй модели была добавлены переменная AREA, при попытке увеличить точность, результат был, но не существенный, точность была равна 0.06, но если смотреть на это в процентном соотношении, точность была увеличена почти в 6000%, с одной стороны цифра огромная, но если смотреть на ситуацию в целом все не так радужно и прекрасною

**Выводы**

Результаты анализа данных о преступлениях в Лос-Анджелесе отражают вызовы и ограничения в построении точной модели прогнозирования места преступления. Несмотря на применение различных переменных, точность моделей варьировалась в диапазоне от 0.001 до 0.06, что подчеркивает сложность создания высокоточной модели.

Модель, основанная на переменных AREA и Vict.Age, демонстрировала наилучшие результаты. Однако, эти переменные оказались ограниченными для построения регрессионной модели из-за ограниченного объема данных, что сказалось на точности прогнозов.

Анализ указывает на необходимость дополнительных исследований и подходов для улучшения модели. Использование более сложных моделей, добавление других переменных и увеличение объема данных могут сыграть ключевую роль в улучшении точности прогнозов мест преступлений.

Обсуждая проведенный анализ, гипотеза о зависимости возраста жертвы от места преступления не получила подтверждения из-за сложности создания достаточно точной модели. Добавление переменной AREA позволило улучшить точность, но незначительно. Важно отметить, что увеличение точности модели на 6000% не означает существенного улучшения, поскольку исходная точность была на низком уровне.

Этот анализ выделяет важность дальнейших исследований и разработки более сложных моделей для более точного прогнозирования мест преступлений в городе.

**Библиография**

1. Ссылка на исходные кода и все данные о датасете ([ronmavoitenko/AD (github.com)](https://github.com/ronmavoitenko/AD))
2. Датасет о преступлениях ([Crime Data from 2020 to Present - Catalog](https://catalog.data.gov/dataset/crime-data-from-2020-to-present))
3. Документация по R ([Crime Data from 2020 to Present - Catalog](https://catalog.data.gov/dataset/crime-data-from-2020-to-present))
4. Построение регрессивной модели, изучение ([RPubs - Построение модели линейной регрессии](https://rpubs.com/vasorokina/472416))
5. Линейные модели ([Открытый курс машинного обучения. Тема 4. Линейные модели классификации и регрессии / Хабр (habr.com)](https://habr.com/ru/companies/ods/articles/323890/))
6. Работа с CSV файлами ([Как импортировать файлы CSV в R (шаг за шагом) (codecamp.ru)](https://www.codecamp.ru/blog/import-csv-into-r/))