Анализ данных:
с чего начинать,
что делать нужно, что можно,
а что – нельзя

ЛЭШ, 6 августа 2023

Ход исследования и данные

Этапы исследования

- Исследовательский вопрос
- Формулировка гипотезы
  - «Гипотеза – утверждение, помогающее ответить на исследовательский вопрос или на часть его (предположение об ответе)»
  - Механизм!
- Данные
  - Предварительный анализ:
    - Почитать документацию
    - Посчитать описательные статистики
    - Поработать с выбросами
    - Подготовить данные (трансформации, нормировка и т.д.)
  - Анализ, оценка моделей
  - Визуализация результатов
  - Интерпретация
  - Выводы и объяснение, что с полученными результатами вообще делать

Какие данные вообще брать?

Очень часто перед началом работы с данными нужно понять, какие данные в принципе нужны

Какие данные вообще брать?

Очень часто перед началом работы с данными нужно понять, какие данные в принципе нужны

- «Разрешение» данных:
  - Что должно быть объектом наблюдения? Страна, регион, группа людей, отдельный человек?
  - Насколько детальные данные нужны? К примеру: продажи отдельных продуктов, узких групп, широких групп или продажи целиком?

Какие данные вообще брать?

Очень часто перед началом работы с данными нужно понять, какие данные в принципе нужны

- «Разрешение» данных:
  - Что должно быть объектом наблюдения? Страна, регион, группа людей, отдельный человек?
  - Насколько детальные данные нужны? К примеру: продажи отдельных продуктов, узких групп, широких групп или продажи целиком?
- Частота данных:
  - Данные могут быть миллисекундными, а могут быть – годовыми
  - Выбор зависит от задачи!

Какие данные вообще брать?

Очень часто перед началом работы с данными нужно понять, какие данные в принципе нужны

- «Разрешение» данных:
  - Что должно быть объектом наблюдения? Страна, регион, группа людей, отдельный человек?
  - Насколько детальные данные нужны? К примеру: продажи отдельных продуктов, узких групп, широких групп или продажи целиком?
- Частота данных:
  - Данные могут быть миллисекундными, а могут быть – годовыми
  - Выбор зависит от задачи!
- Средние за период показатели, суммарные, минимумы, максимумы, на конец периода, на начало периода?

Предварительный анализ: документация

- Читать документацию к данным почти всегда скучно

Сельское и лесное хозяйство

Объем производства продукции сельского хозяйства всеми сельхозпроизводителями (сельхозорганизации, крестьянские (фермерские) хозяйства, индивидуальные предприниматели, хозяйства населения) формируется как объем производства готовой продукции растениеводства и животноводства и изменение стоимости незавершенного производства продукции растениеводства и животноводства по видам деятельности "Растениеводство", "Животноводство". При этом объем производства скота и птицы определяется на уровне объема выращивания скота. Выращивание продуктивного скота и птицы складывается из веса полученного за отчетный период приплода, веса прироста молодняка и привеса скота и птицы (взрослых и молодняка) в результате их откорма и нагула за минусом падежа молодняка скота и скота на откорме. Объем продукции выращивания исчисляется в живом весе и рассчитывается по основным видам продуктивного скота (крупный рогатый скот, свиньи, овцы и козы и др.), а также по всем видам птицы.

Индекс производства продукции сельского хозяйства - относительный показатель, характеризующий изменение объема производства сельскохозяйственной продукции в сравниваемых периодах. Различают индивидуальные и сводные индексы производства. Индивидуальные индексы отражают изменение производства одного продукта и исчисляются как отношение объемов производства данного вида продукции в натуральном выражении в сравниваемых периодах. Сводный индекс производства характеризует совокупные изменения производства всех видов продукции в результате изменения только физического объема производимой продукции.

Индекс производства продукции сельского хозяйства - агрегированный индекс производства продукции растениеводства и животноводства. Для исчисления индекса производства продукции сельского хозяйства к соответствующему периоду предыдущего года используется показатель ее объема в сопоставимых ценах предыдущего года.

Оценка представленных на графике данных по индексам производства с исключением сезонного фактора осуществляется с использованием метода TRAMO-SEATS программы "JDemetra+". При построении модели не учитываются календарные эффекты в связи с особенностями сельскохозяйственного производства. Модель фиксируется в начале года и не меняется на его протяжении.

В качестве исходной информации для проведения сезонной корректировки применяются ежемесячные фактические значения индексов производства продукции сельского хозяйства с 2009 г., рассчитанные в процентах к среднемесячному значению 2020 года.

Данные о поголовье сельскохозяйственных животных, производстве и реализации основных сельскохозяйственных продуктов по всем сельхозпроизводителям определяются: по сельскохозяйственным организациям - на основании сведений форм федерального статистического наблюдения (по субъектам малого предпринимательства - с применением выборочного метода наблюдения); по хозяйствам населения, крестьянским (фермерским) хозяйствам и индивидуальным предпринимателям - по материалам выборочных обследований.

Производство скота и птицы на убой (в живом весе) включает проданные сельхозпроизводителями скот и птицу для забоя на мясо, а также забитые в сельскохозяйственных организациях, крестьянских (фермерских) хозяйствах, у индивидуальных предпринимателей, в хозяйствах населения.

Производство молока характеризуется фактически надоенным сырым коровьим, козьим, овечьим, кобыльим и буйволиным молоком. Молоко, высосанное молодняком при подсосном его содержании, в продукцию не включается.

Производство яиц включает их сбор от всех видов сельскохозяйственной птицы, в том числе и яйца, пошедшие на воспроизводство птицы (инкубация и др.).

Информация об обеспеченности скота кормами (публикуется в докладах № 1-4, 9-12).

Сельское и лесное хозяйство

Объем производства продукции сельского хозяйства всеми сельхозпроизводителями (сельхозорганизации, крестьянские (фермерские) хозяйства, индивидуальные предприниматели, хозяйства населения) формируется как объем производства готовой продукции растениеводства и животноводства и изменение стоимости незавершенного производства продукции растениеводства и животноводства по видам деятельности "Растениеводство", "Животноводство". При этом **объем производства скота и птицы определяется на уровне объема выращивания скота**. Выращивание продуктивного скота и птицы складывается из **веса полученного за отчетный период приплода, веса прироста молодняка и привеса скота и птицы (взрослых и молодняка) в результате их откорма и нагула за минусом падежа молодняка скота и скота на откорме**. Объем продукции выращивания исчисляется в живом весе и рассчитывается по основным видам продуктивного скота (крупный рогатый скот, свиньи, овцы и козы и др.), а также по всем видам птицы.

Индекс производства продукции сельского хозяйства - относительный показатель, характеризующий изменение объема производства сельскохозяйственной продукции в сравниваемых периодах. Различают индивидуальные и сводные индексы производства. Индивидуальные индексы отражают изменение производства одного продукта и исчисляются как отношение объемов производства данного вида продукции в натуральном выражении в сравниваемых периодах. Сводный индекс производства характеризует совокупные изменения производства всех видов продукции в результате изменения только физического объема производимой продукции.

Индекс производства продукции сельского хозяйства - агрегированный индекс производства продукции растениеводства и животноводства. Для исчисления индекса производства продукции сельского хозяйства к соответствующему периоду предыдущего года используется показатель ее объема в сопоставимых ценах предыдущего года.

Оценка представленных на графике данных по индексам производства с исключением сезонного фактора осуществляется с использованием метода TRAMO-SEATS программы "JDemetra+". При построении модели не учитываются календарные эффекты в связи с особенностями сельскохозяйственного производства. Модель фиксируется в начале года и не меняется на его протяжении.

В качестве исходной информации для проведения сезонной корректировки применяются ежемесячные фактические значения индексов производства продукции сельского хозяйства с 2009 г., рассчитанные в процентах к среднемесячному значению 2020 года.

Данные о поголовье сельскохозяйственных животных, производстве и реализации основных сельскохозяйственных продуктов по всем сельхозпроизводителям определяются: по сельскохозяйственным организациям - на основании сведений форм федерального статистического наблюдения (по субъектам малого предпринимательства - с применением выборочного метода наблюдения); по хозяйствам населения, **крестьянским (фермерским) хозяйствам и индивидуальным предпринимателям - по материалам выборочных обследований**.

Производство скота и птицы на убой (в живом весе) включает проданные сельхозпроизводителями скот и птицу для забоя на мясо, а также забитые в сельскохозяйственных организациях, крестьянских (фермерских) хозяйствах, у индивидуальных предпринимателей, в хозяйствах населения.

Производство молока характеризуется фактически надоенным сырым коровьим, козьим, овечьим, кобыльим и буйволиным молоком. Молоко, высосанное молодняком при подсосном его содержании, в продукцию не включается.

Производство яиц включает их сбор от всех видов сельскохозяйственной птицы, в том числе и яйца, пошедшие на воспроизводство птицы (инкубация и др.).

Информация об обеспеченности скота кормами (публикуется в докладах № 1-4, 9-12).

Предварительный анализ: документация

- Читать документацию к данным почти всегда скучно
- И часто ещё и долго
  - Методологические пояснения к СЭПР Росстата: 36 страниц

Предварительный анализ: документация

- Читать документацию к данным почти всегда скучно
- И часто ещё и долго
  - Методологические пояснения к СЭПР Росстата: 36 страниц
- Или даже очень долго
  - Codebooks к последней опубликованной волне RLMS: 298 страниц для данных по домохозяйствам + 87 страниц для данных по индивидам

Предварительный анализ: документация

- Читать документацию к данным почти всегда скучно
- И часто ещё и долго
  - Методологические пояснения к СЭПР Росстата: 36 страниц
- Или даже очень долго
  - Codebooks к последней опубликованной волне RLMS: 298 страниц для данных по домохозяйствам + 87 страниц для данных по индивидам
- Но! Если документацию совсем не читать – может получиться ещё хуже

Почему важно читать документацию

- В любых официальных источниках можно найти темпы роста ВВП практически любой более-менее крупной страны. К примеру, Германии!

Почему важно читать документацию

- О ужас, в 2015 году произошла катастрофа!

(примерно такими темпами, по 15% в год, падал ВВП США в первые годы Великой Депрессии)

Почему важно читать документацию

- Или нет?

Почему важно читать документацию

- А всё дело в том, что были взяты разные цены!
- А евро ослаб с 1.34 доллара за евро в июне 2014 до 1.02 в июне 2015

Предварительный анализ: описательные статистики и выбросы

- Описательные статистики (хотя бы: среднее, стандартное отклонение, минимум, максимум) позволяют понять, что вообще происходит в ваших данных
- В том числе – они позволяют найти выбросы и ошибки кодировки!
- Выбросы, особенно в опросных данных, встречаются очень часто

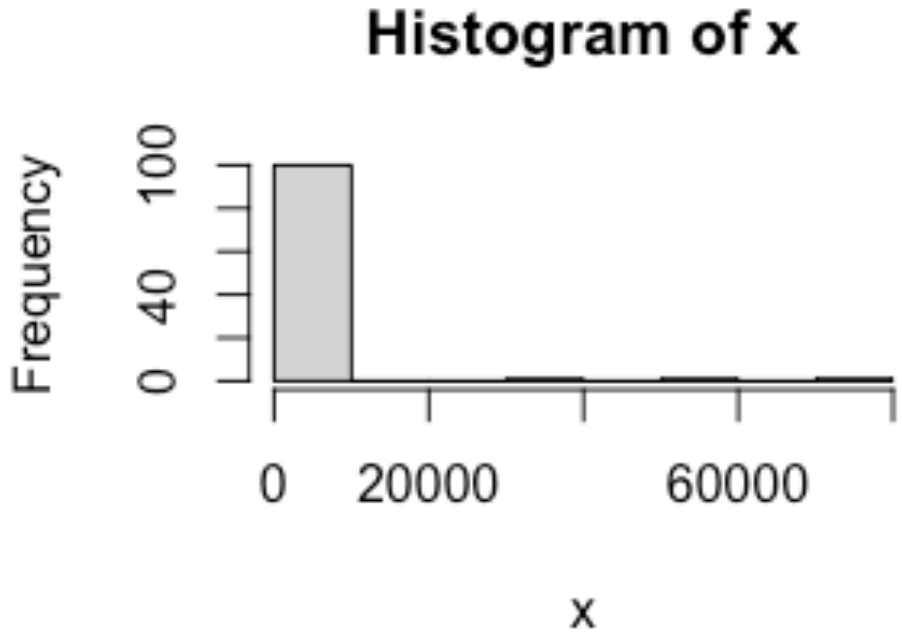
Предварительный анализ: описательные статистики и выбросы

- Пример: описательные статистики переменной «Ежемесячный доход, тыс. руб.»

```
> summary(x)
   Min.   1st Qu.   Median     Mean     3rd Qu.      Max.
  30.22    43.68     50.67     1748.25    58.13     75000.00
```

Предварительный анализ: описательные статистики и выбросы

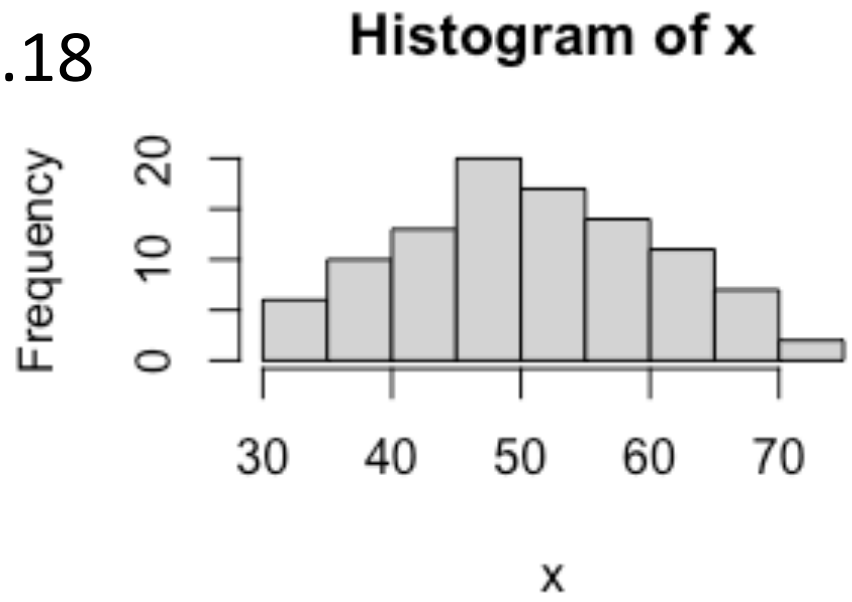
- Пример: описательные статистики переменной «Ежемесячный доход, тыс. руб.»



Предварительный анализ: описательные статистики и выбросы

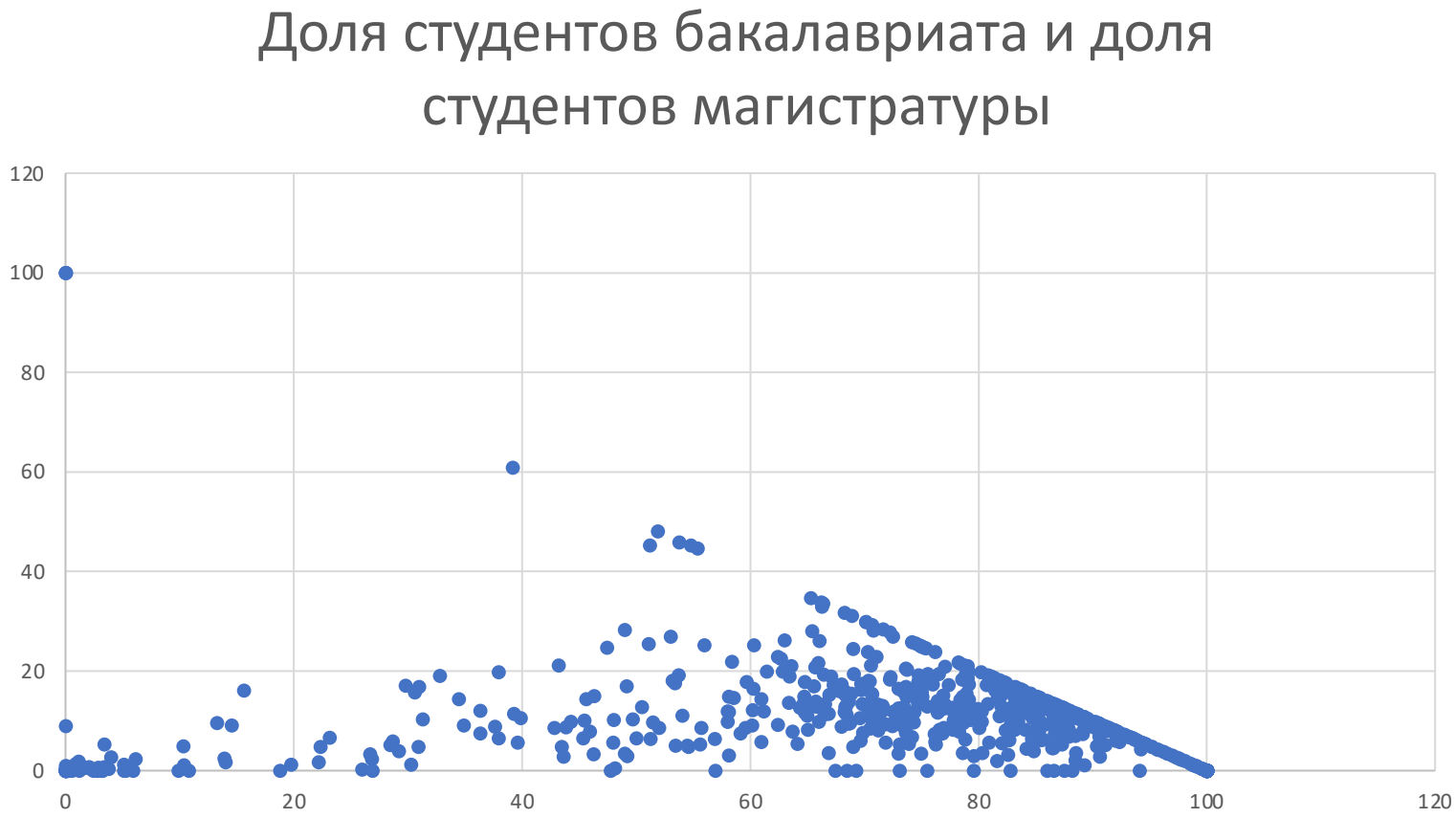
- Пример: описательные статистики переменной «Ежемесячный доход, тыс. руб.» без выбросов

```
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  30.22   43.58   50.44   50.70   57.07   74.18
```



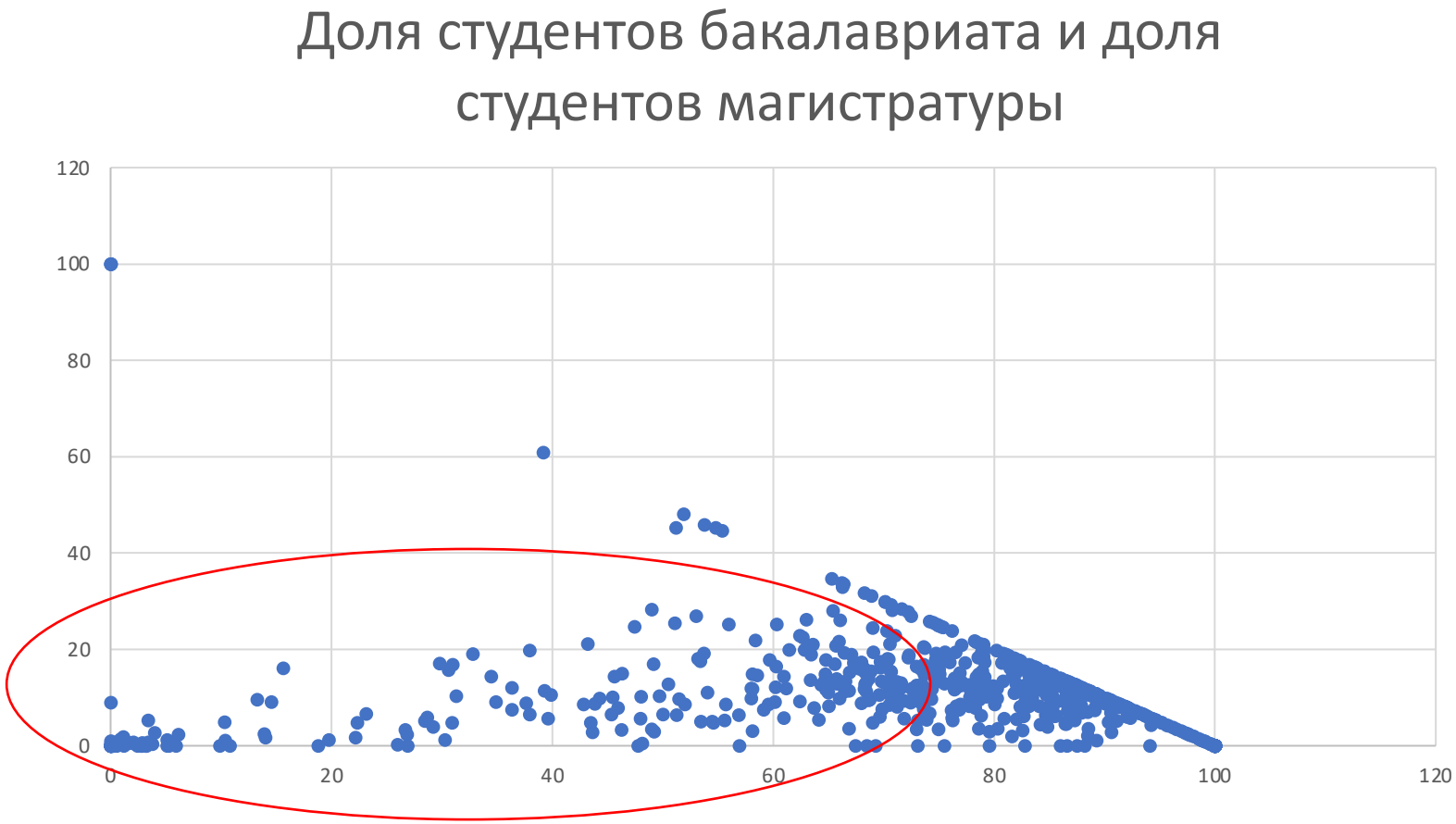
Предварительный анализ: описательные статистики и выбросы

- Пример 2: ВУЗы



Предварительный анализ: описательные статистики и выбросы

- Пример 2: ВУЗы

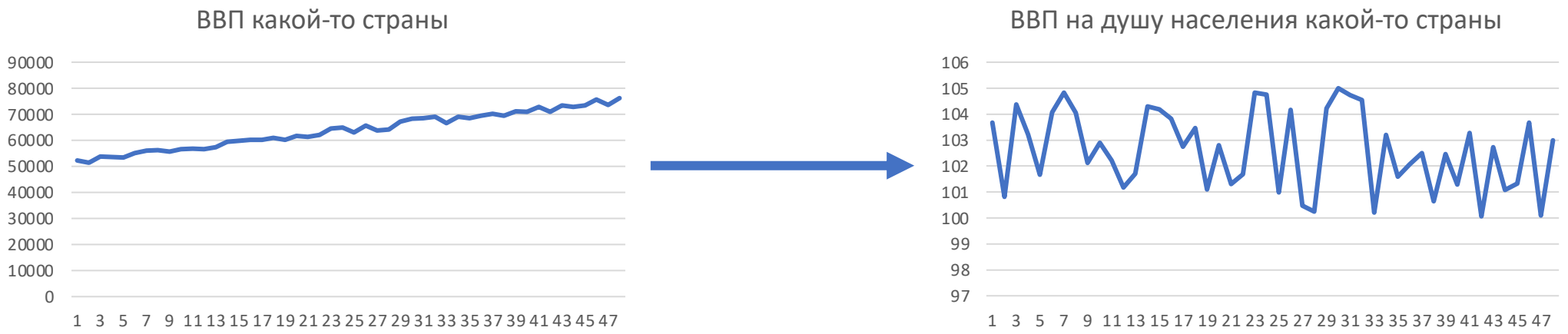


Подготовка данных

- Базовые трансформации данных

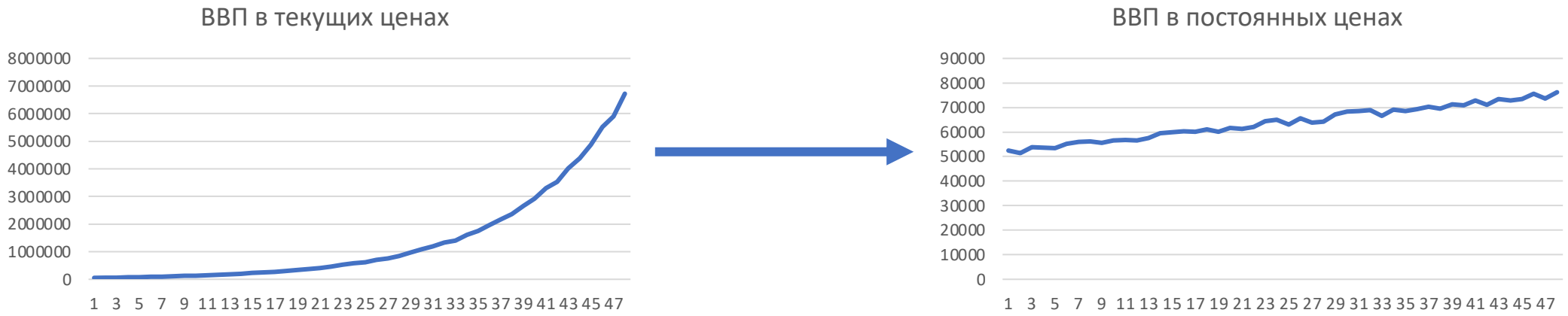
Подготовка данных

- Базовые трансформации данных
  - Учесть размер
    - Не ВВП стран / ВРП регионов, а ВВП / ВРП на душу населения



Подготовка данных

- Базовые трансформации данных
  - Учесть размер
    - Не ВВП стран / ВРП регионов, а ВВП / ВРП на душу населения
  - Привести к одинаковым ценам
    - Каким? Нацвалюта / доллары; ППС; текущие цены или постоянные цены?



Подготовка данных

- Базовые трансформации данных
  - Учесть размер
    - Не ВВП стран / ВРП регионов, а ВВП / ВРП на душу населения
  - Привести к одинаковым ценам
    - Каким? Нацвалюта / доллары; ППС; текущие цены или постоянные цены?
  - Календарные преобразования

Подготовка данных

- Базовые трансформации данных
  - Учесть размер
    - Не ВВП стран / ВРП регионов, а ВВП / ВРП на душу населения
  - Привести к одинаковым ценам
    - Каким? Нацвалюта / доллары; ППС; текущие цены или постоянные цены?
  - Календарные преобразования
  - Математические преобразования (к примеру, переход к темпам роста)

Подготовка данных

- Базовые трансформации данных
  - Учесть размер
    - Не ВВП стран / ВРП регионов, а ВВП / ВРП на душу населения
  - Привести к одинаковым ценам
    - Каким? Нацвалюта / доллары; ППС; текущие цены или постоянные цены?
  - Календарные преобразования
  - Математические преобразования (к примеру, переход к темпам роста)
- Корректировка сезонности и эффектов плавающих праздников

Подготовка данных

- Базовые трансформации данных
  - Учесть размер
    - Не ВВП стран / ВРП регионов, а ВВП / ВРП на душу населения
  - Привести к одинаковым ценам
    - Каким? Нацвалюта / доллары; ППС; текущие цены или постоянные цены?
  - Календарные преобразования
  - Математические преобразования (к примеру, переход к темпам роста)
- Корректировка сезонности и эффектов плавающих праздников
- Скользящее среднее, заполнение пропусков и прочее

Анализ, модели, результаты

- Очень важно понимать, что в принципе можно изучать по данным, а что – нельзя
- К примеру: На основе данных мониторинга ВУЗов может возникнуть идея измерить качество образования в ВУЗах

Анализ, модели, результаты

- Очень важно понимать, что в принципе можно изучать по данным, а что – нельзя
- К примеру: На основе данных мониторинга ВУЗов может возникнуть идея измерить качество образования в ВУЗах

- Как можно измерить качество образования?

Анализ, модели, результаты

- Очень важно понимать, что в принципе можно изучать по данным, а что – нельзя
- К примеру: На основе данных мониторинга ВУЗов может возникнуть идея измерить качество образования в ВУЗах

- Как можно измерить качество образования?
  - Востребованность выпускников среди работодателей (доходы выпускников, время поиска работы, доля выпускников, работающих по специальности)
  - Доля выпускников, получающих в дальнейшем учёную степень

Анализ, модели, результаты

- Очень важно понимать, что в принципе можно изучать по данным, а что – нельзя
- К примеру: На основе данных мониторинга ВУЗов может возникнуть идея измерить качество образования в ВУЗах

- Как можно измерить качество образования?
  - Востребованность выпускников среди работодателей (доходы выпускников, время поиска работы, доля выпускников, работающих по специальности)
  - Доля выпускников, получающих в дальнейшем учёную степень
- Этих данных в мониторинге просто нет!
- Есть данные по качеству приёма – но это уже другое исследование

Где брать данные

Макроданные по России: Росстат и ЦБ

- Росстат – лучший источник данных по России в целом и регионам
  - Всё основное в одном месте:
    - https://rosstat.gov.ru/folder/10705
  - Оперативные данные: СЭПР
    - https://rosstat.gov.ru/compendium/document/50801
  - Много детальных тематических данных – публикации
    - https://rosstat.gov.ru/compendium
  - Региональная статистика
    - https://rosstat.gov.ru/folder/11109/document/13259
  - Есть BI-система, но работает она медленно (данные те же самые)
    - http://bi.gks.ru/biportal/contourbi.jsp

Макроданные по России: Росстат и ЦБ

- Много информации по банковской системе есть у ЦБ:
  - https://www.cbr.ru/statistics/
  - Макростатистика:
    - https://www.cbr.ru/statistics/macro_itm/
  - Характеристики банковского сектора:
    - https://www.cbr.ru/statistics/bank_sector/
    - В том числе в разрезе отдельных банков: https://www.cbr.ru/statistics/bank_sector/pdko_sub/
  - Состояние финансовых рынков:
    - https://www.cbr.ru/statistics/finr/

Микроданные по России: РМЭЗ

- РМЭЗ (RLMS) – ежегодный репрезентативный опрос населения (единица наблюдения – человек) по очень широкому кругу вопросов
  - https://www.hse.ru/rlms
- Опрос ежегодные, данные доступны за период с 1994 по 2021
  - https://www.hse.ru/rlms/spss
- Работать с данными достаточно сложно. Описание переменных доступно здесь:
  - https://www.hse.ru/rlms/code

Макроданные по другим странам

- Данные Всемирного Банка
  - https://databank.worldbank.org
- Данные ОЭСР
  - https://data.oecd.org
- Данные МВФ
  - https://www.imf.org/en/Data
- FRED
  - https://fred.stlouisfed.org

Макроданные по другим странам

- Более тематические данные можно найти у профильных международных организаций
  - Здравоохранение и медицина – ВОЗ
    - https://www.who.int/data
  - Продовольствие – FAO
    - https://www.fao.org/faostat/en/#home
  - Труд – ILO
    - https://ilostat.ilo.org

Приложение

Вопрос – гипотеза – механизм

Исследовательский вопрос должен быть интересным и нетривиальным

Гипотеза – утверждение, помогающие ответить на исследовательский вопрос

Механизм объясняет, почему гипотеза работает

Вопрос – гипотеза – механизм

Исследовательский вопрос должен быть интересным и нетривиальным

- В какой ВУЗ поступает больше всего победителей олимпиад – тривиальный вопрос
- Какие характеристики ВУЗа привлекают в него победителей олимпиад - нетривиальный

Гипотеза – утверждение, помогающие ответить на исследовательский вопрос

Механизм объясняет, почему гипотеза работает

Вопрос – гипотеза – механизм

Исследовательский вопрос должен быть интересным и нетривиальным

- В какой ВУЗ поступает больше всего победителей олимпиад – тривиальный вопрос
- Какие характеристики ВУЗа привлекают в него победителей олимпиад - нетривиальный

Гипотеза – утверждение, помогающие ответить на исследовательский вопрос

- Победители олимпиад охотнее поступают в ВУЗы, где преподавателям больше платят
- Победители олимпиад охотнее поступают в ВУЗы, где преподаватели больше публикуются в хороших научных журналах

Механизм объясняет, почему гипотеза работает

Вопрос – гипотеза – механизм

Исследовательский вопрос должен быть интересным и нетривиальным

- В какой ВУЗ поступает больше всего победителей олимпиад – тривиальный вопрос
- Какие характеристики ВУЗа привлекают в него победителей олимпиад - нетривиальный

Гипотеза – утверждение, помогающие ответить на исследовательский вопрос

- Победители олимпиад охотнее поступают в ВУЗы, где преподавателям больше платят
- Победители олимпиад охотнее поступают в ВУЗы, где преподаватели больше публикуются в хороших научных журналах

Механизм объясняет, почему гипотеза работает

- В ВУЗах с более высокими зарплатами работают более квалифицированные преподаватели, а из ВУЗов с низкими зарплатами они уходят
- Преподаватели со статьями в хороших научных журналах больше знают и способны большему научить

Примеры исследований на данных

Телематические данные и моделирование аварийности автомобилей

- По мотивам *Stankevich I. et al. Usage-based vehicle insurance: Driving style factors of accident probability and severity //Journal of Transportation Safety & Security. – 2021. – С. 1-22*
- Почему это важно и интересно?
  - Разные люди ездят по-разному
  - Как стиль вождения влияет на аварийность?
  - Более резкое и опасное вождение должно повышать вероятность попадания в аварию

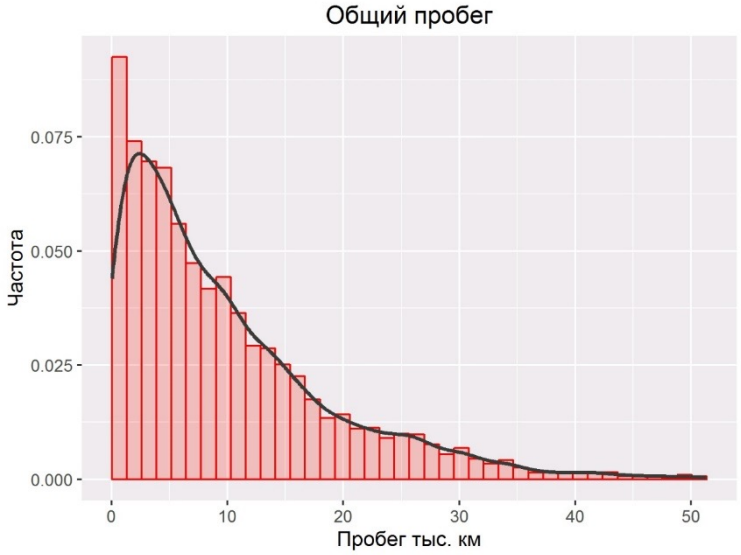
Телематические данные и моделирование аварийности автомобилей

- В автомобили устанавливаются устройства, фиксирующие координаты, скорость движения и ускорения автомобиля
- Идея: на основе полученных данных оценить качество вождения
- Аккуратные водители получают скидки на страховку
- Неаккуратные платят за страховку больше
- Помимо этого, телематика позволяет контролировать водителей корпоративных автопарков, искать нелегальных таксистов и много чего еще

Телематические данные и моделирование аварийности автомобилей

- Исходные данные – приходящие с интервалом от 0.1 до 5 секунд пакеты с координатами, ускорениями по 3 осям, скоростью и служебной информацией
- Задача 1: сформировать на основе этих данных переменные, которые будут:
  - Интерпретируемыми
  - Хорошо описывать стиль вождения

Анализ данных: пробег



С точки зрения распределения машин в выборке по общему и среднему пробегу, наблюдается одномодальное распределение с тяжелым правым хвостом.

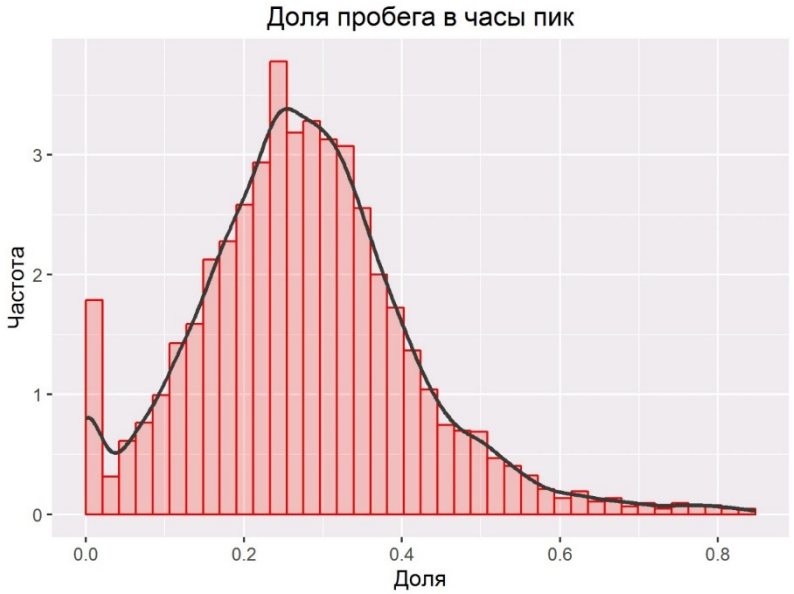
То есть наибольшее количество водителей сосредоточено вокруг типичного значения (для среднедневного пробега – 40 км).

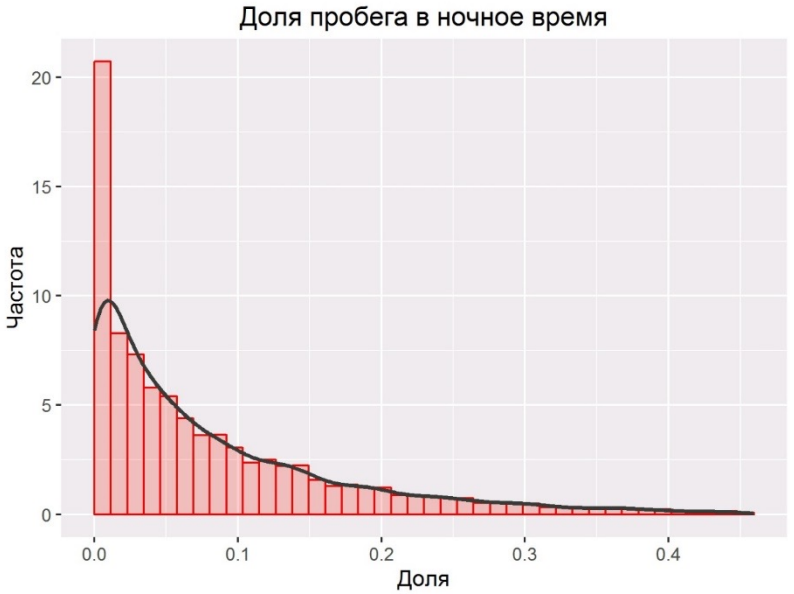
Поведение остальных – отклонения от этого типичного значения (для среднедневного пробега максимум превышает 200 км).

Анализ данных: структура пробега

Распределение машин по доли пробега в часы пик имеет два локальных максимума: нулевые значения (без поездок в часы пик) и порядка 25% (глобальный максимум). То есть присутствуют две группы, одна из которых вообще не ездит (порядка 10% от всех) , а другие ведут себя типично вокруг значения в 25%.

Доля пробега в ночное время имеет наиболее типичное значение в нуле. Остальные наблюдения – те, кто ездит ночью, чьи значения достигают 50%.

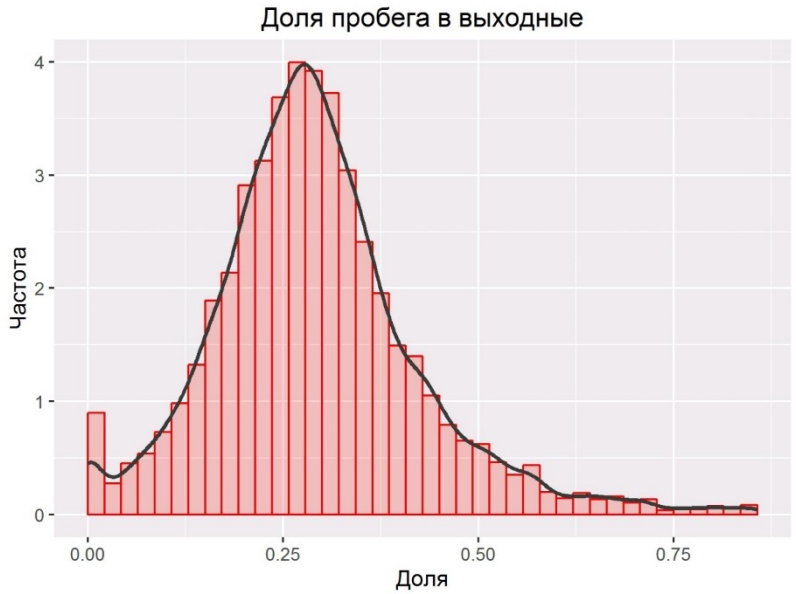


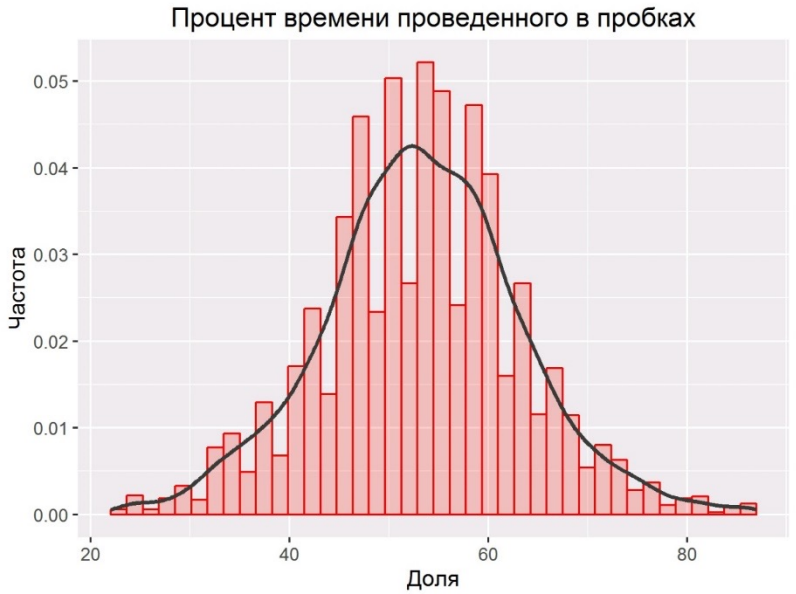


Анализ данных: структура пробега

Распределение машин по доли пробега в выходные имеет характерный пик на отметке в 30%. Вокруг этого пика распределение, в целом, симметричное, плюс небольшой правый хвост, достигающий значений более 80%.

Процент времени проведенного в пробках имеет небольшой локальный максимум в нуле и глобальный максимум в районе 50%. Вокруг глобального максимума распределение симметричное.





Анализ данных: частота ускорений

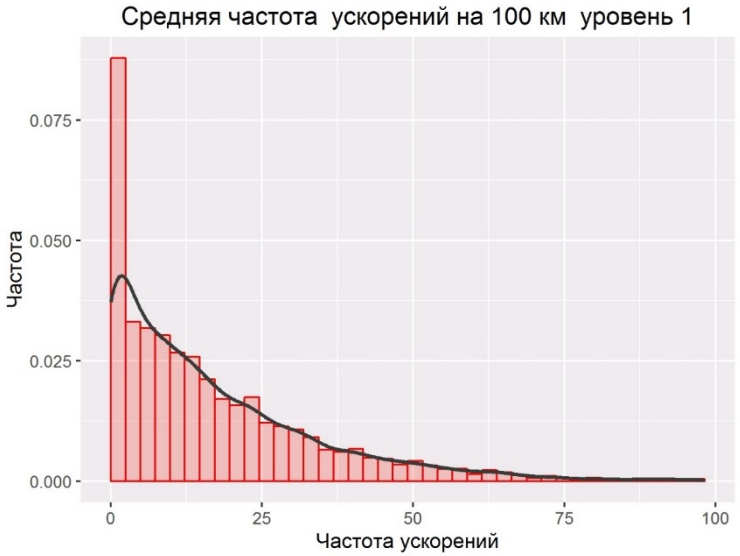
Пороги ускорений:

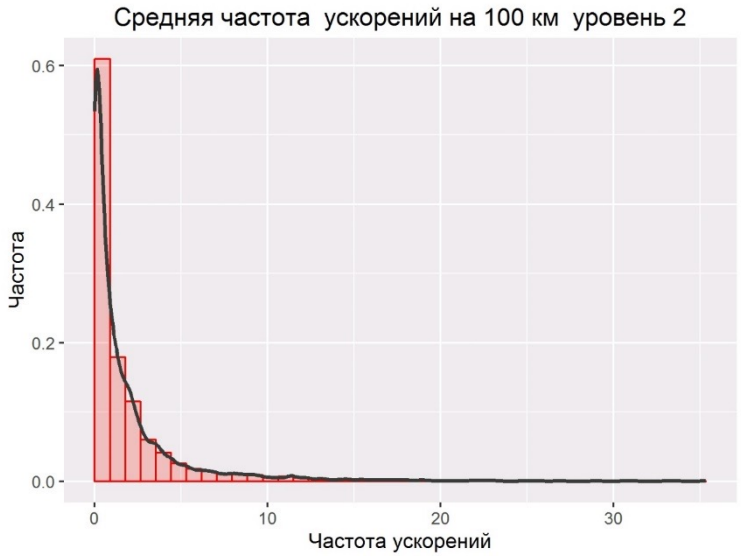
1 уровень – 0.2g (+35 км/ч за 5 сек)

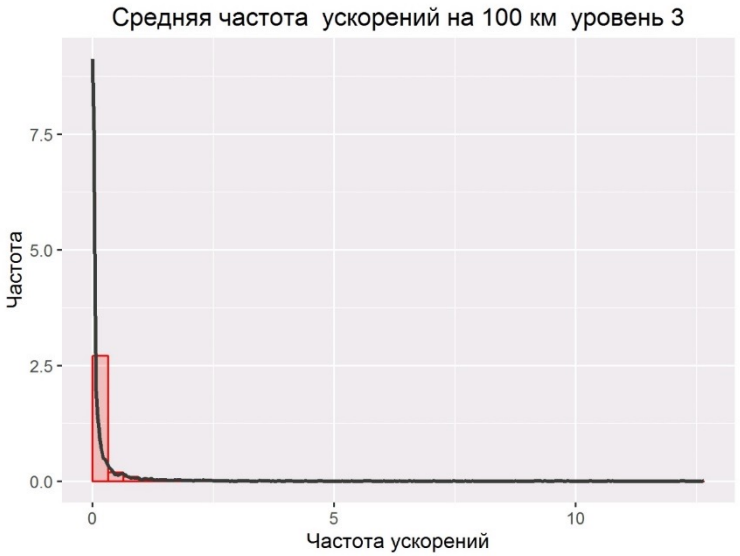
2 уровень – 0.3g (+53 км/ч за 5 сек)

3 уровень – 0.4g (+70 км/ч за 5 сек)

Ускорение 1 уровня совершаются почти в 10 раз чаще, чем ускорения 2 уровня, и в 100 раз чаще, чем ускорения 3 уровня.







Анализ данных: частота торможений

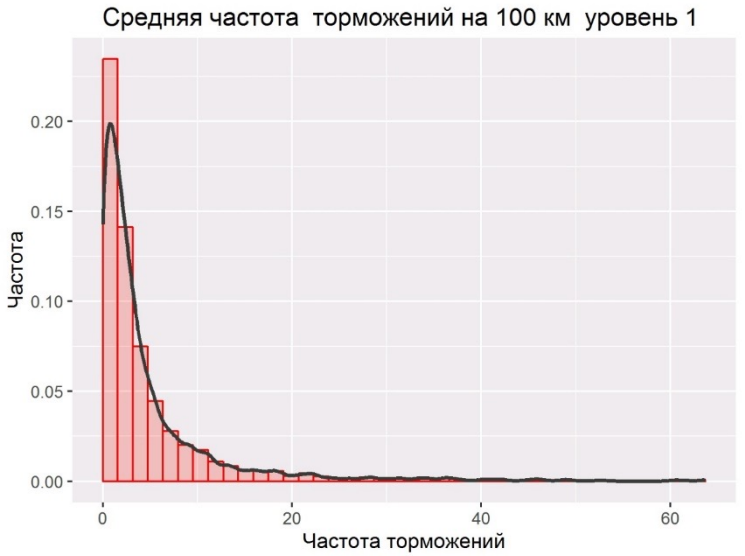
Пороги ускорений:

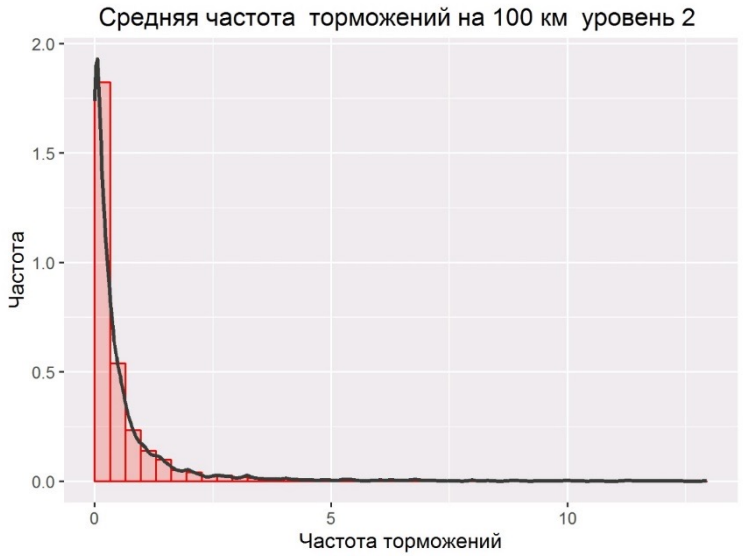
1 уровень – 0.3g (-53 км/ч за 5 сек)

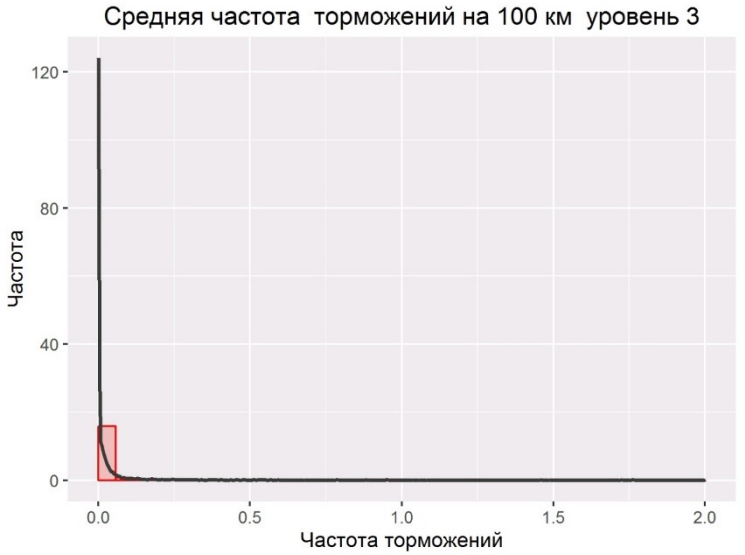
2 уровень – 0.4g (-70 км/ч за 5 сек)

3 уровень – 0.6g (-100 км/ч за 5 сек)

Торможения 1 уровня совершаются почти в 7 раз чаще, чем торможения 2 уровня, и в 50 раз чаще, чем торможения 3 уровня.







Анализ данных: частота боковых ускорений

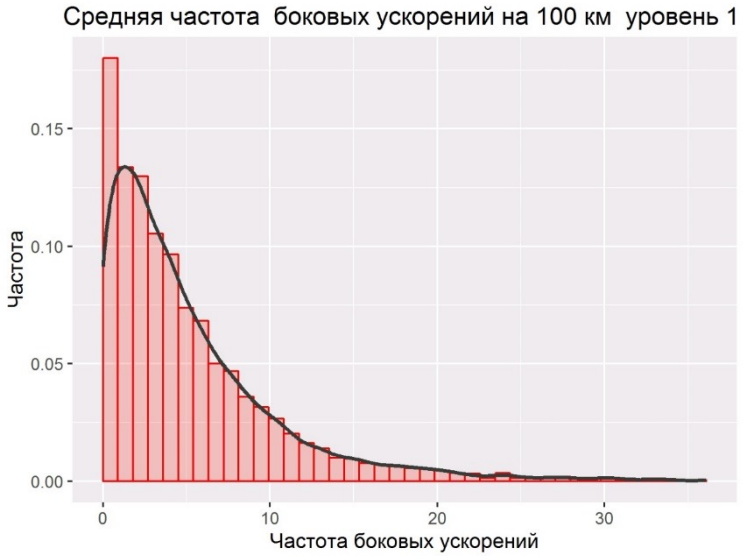
Пороги ускорений:

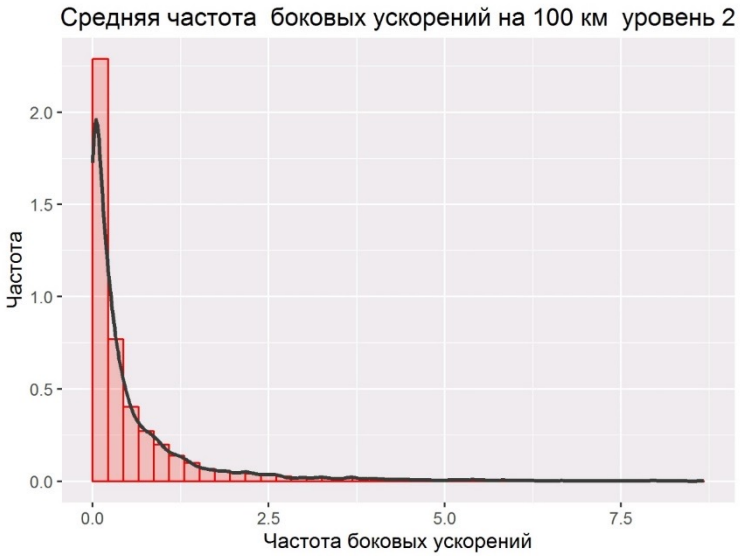
1 уровень – 0.3g (+53 км/ч за 5 сек)

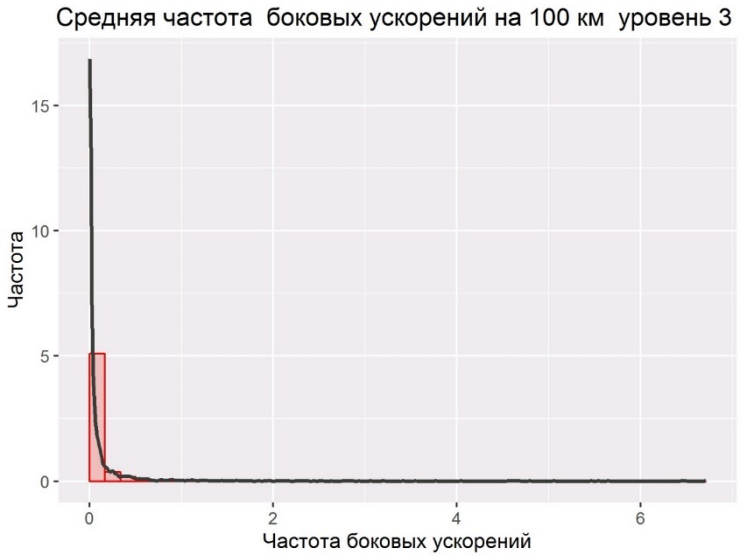
2 уровень – 0.4g (+70 км/ч за 5 сек)

3 уровень – 0.5g (+85 км/ч за 5 сек)

Боковые ускорения 1 уровня совершаются почти в 7 раз чаще, чем боковые ускорений 2 уровня, и в 50 раз чаще, чем боковые ускорения 3 уровня.







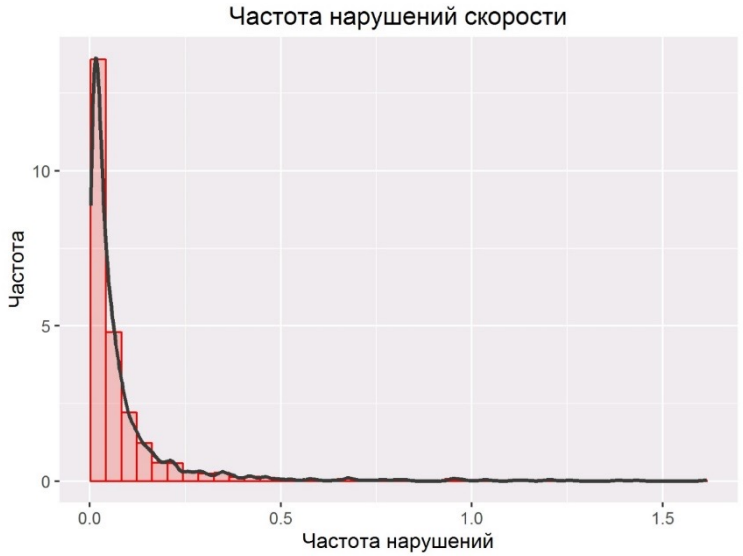
Анализ данных: сезонность в данных по ускорениям

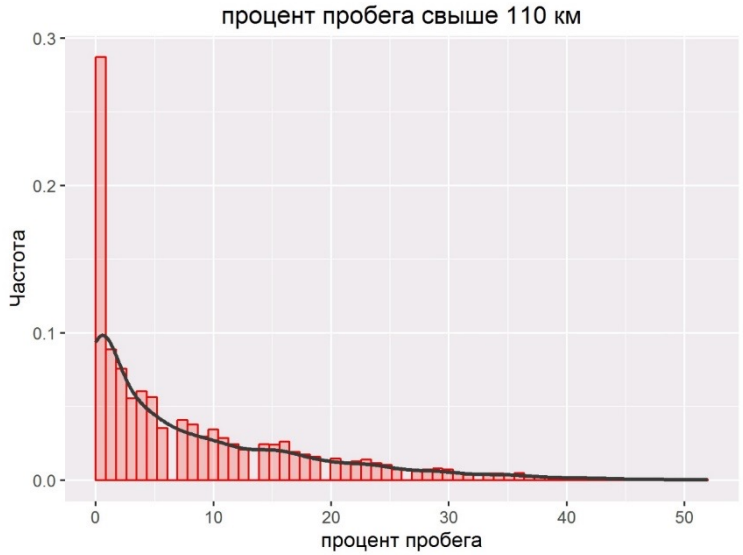
Средняя частота ускорений 1 уровня

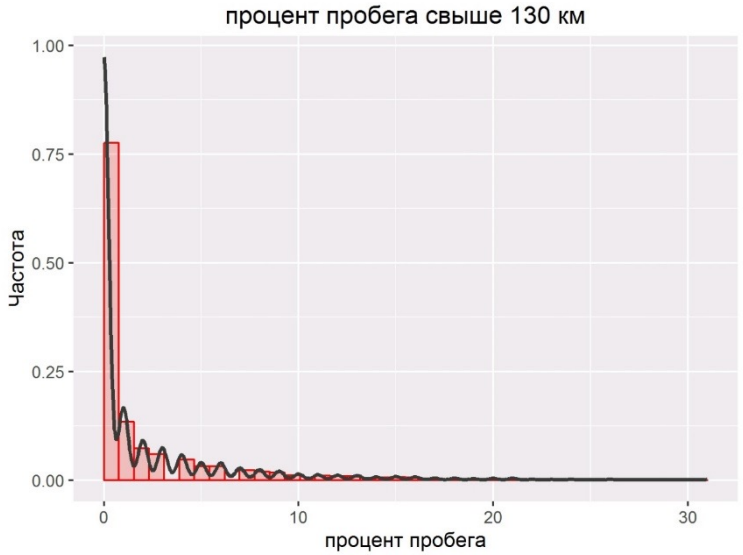
Средняя частота боковых ускорений 2 уровня

Анализ данных: превышения скорости

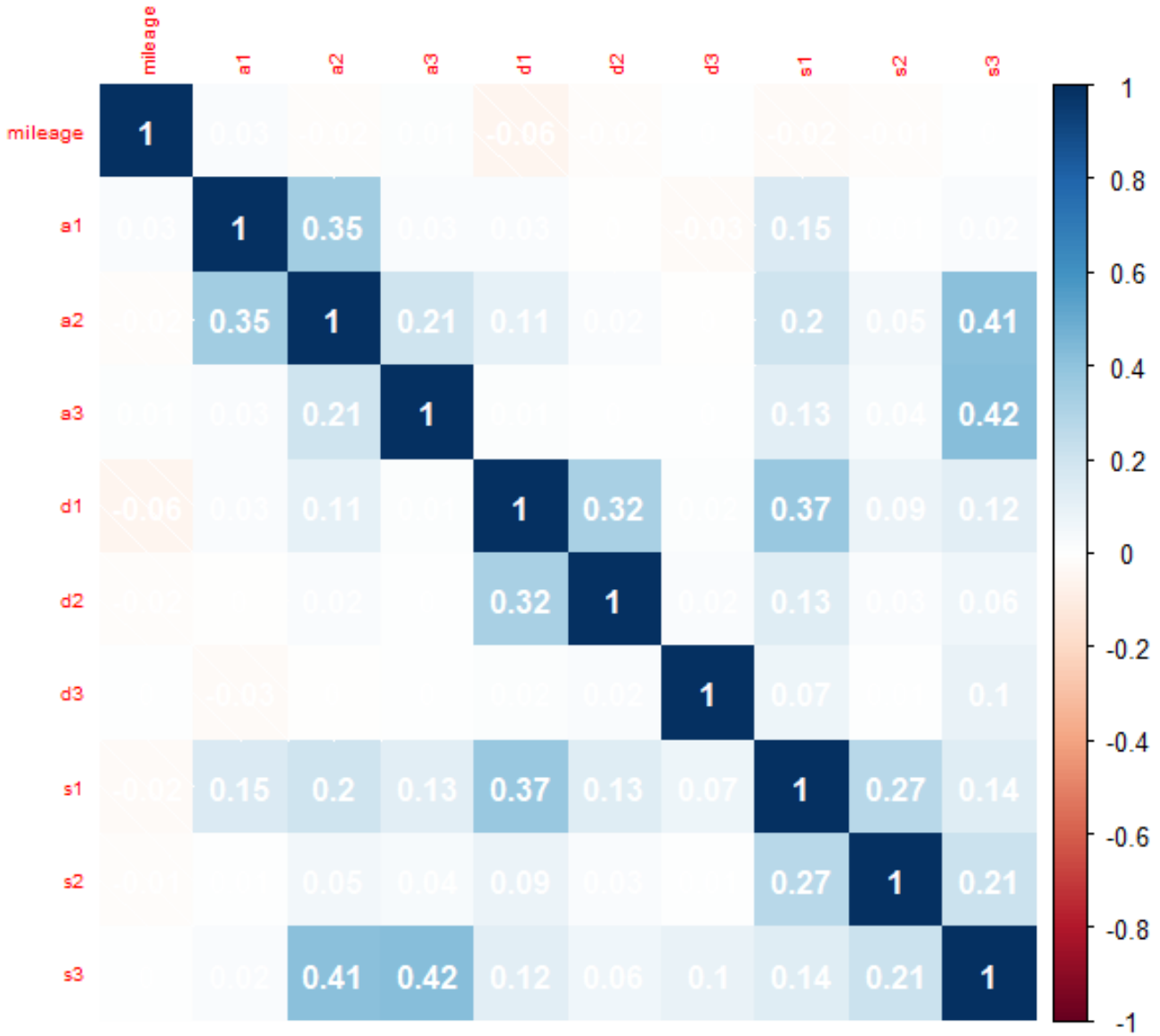
Проценты пробега на скорости свыше 110 и 130 км/ч и частота нарушений скорости имеют пики в нулевых значения и затухающие правые хвосты.







Разведочный анализ: как связаны переменные



Модель вероятности факта аварии

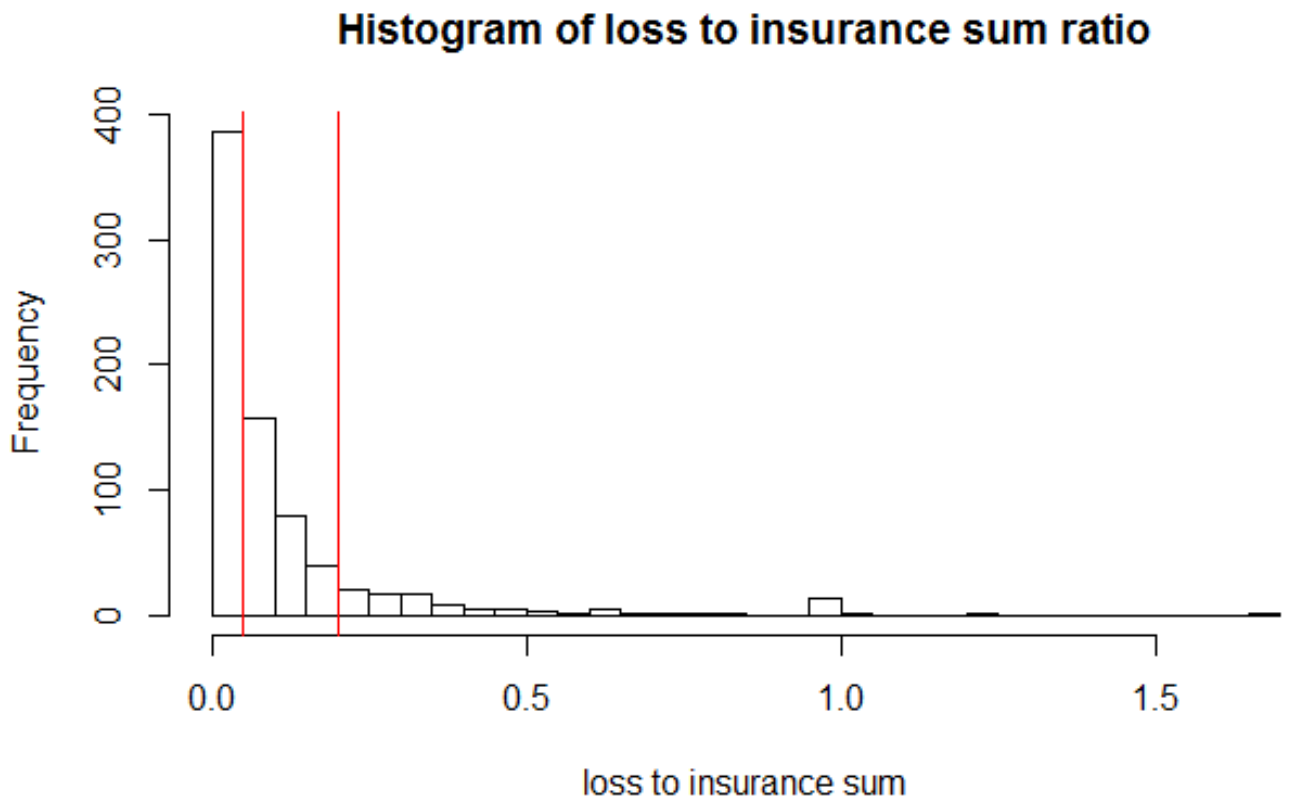
	P-value	Влияние
Общий пробег	0.000	+
Число ускорений 1 уровня	0.007	+
Средняя скорость	0.000	-
Максимальная ночная скорость	0.000	+
Максимальная утренняя скорость	0.015	+

Типичный водитель с высокой вероятностью попасть в аварию любого типа:

- имеет большой общий пробег,
- совершает большое количество ускорений 1 уровня
- вынужден ездить с небольшой средней скоростью,
- ездит ночью и утром с очень высокой скоростью.

Аварии бывают разные!

Отношение потерь к страховой сумме



Модель вероятности слабой аварии

	P-value	Влияние
Общий пробег	0.000	+
Средняя скорость	0.007	-
Максимальная утренняя скорость	0.005	+
Ускорения 1 уровня	0.020	+
Боковые ускорения 1 уровня	0.000	-

Типичный водитель с высокой вероятностью попасть в слабую аварию:

- имеет большой общий пробег,
- вынужден ездить с небольшой средней скоростью,
- ездит с высокой скоростью утром,
- часто совершает ускорения 1 уровня, но не использует боковые ускорения 1 уровня.

Модель вероятности средней аварии

	P-value	Влияние
Общий пробег	0.000	+
Пробег в ночное время	0.059	+
Максимальная ночная скорость	0.026	+
Число ускорений 1 уровня	0.027	+

Типичный водитель с высокой вероятностью попасть в среднюю аварию:

- имеет большой общий пробег,
- имеет большой пробег ночью,
- в ночное время ездит с большой скоростью,
- совершает большое количество ускорений первого уровня.

Модель вероятности сильной аварии

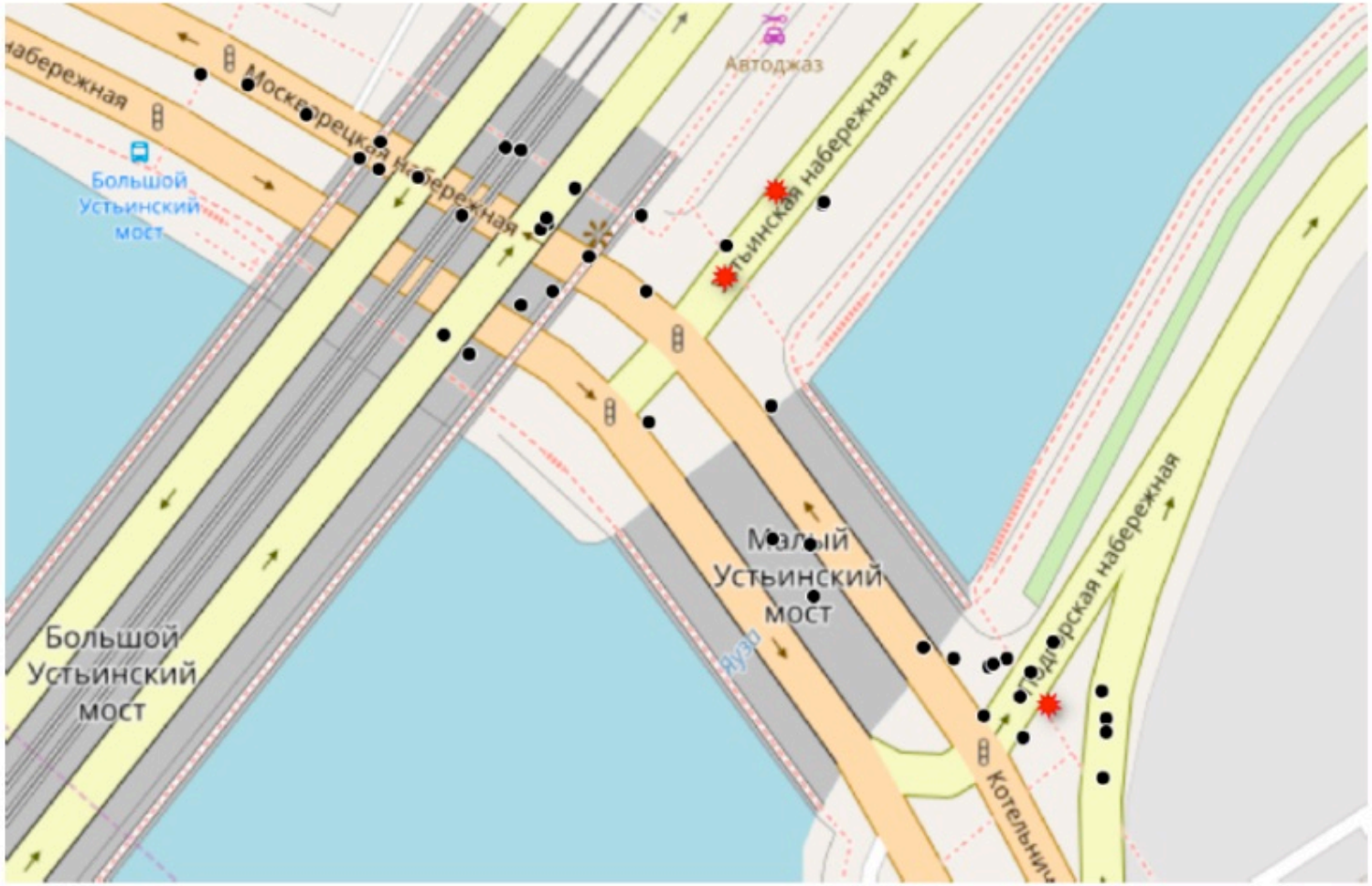
	P-value	Влияние
Максимальная вечерняя скорость	0.004	+
Максимальная ночная скорость	0.027	+
Боковые ускорения 1 уровня	0.023	+
Превышения скорости 3 уровня	0.010	+

Типичный водитель с высокой вероятностью попасть в сильную аварию:

- вечером и ночью ездит с большой скоростью,
- часто совершает боковые ускорения 1 уровня,
- регулярно превышает разрешенную скорость свыше 60 км/ч.

Что можно сделать дальше? Новые переменные!

Изучить поведение водителей в окрестности светофоров:



Что можно сделать дальше? Новые переменные!

Table 11-b

	Full model with clusters	
	Coef	P-value
min_x_more_mean	-0.0032	0.450
max_x_more_mean	0.006	0.001 ***
min_y_more_mean	-0.0038	0.580
max_y_more_mean	0.0017	0.643
min_z_more_mean	0.0009	0.711
max_z_more_mean	-0.0028	0.231
min_x_less_mean	-0.0146	0.004 **
max_x_less_mean	0.0082	0.102
min_y_less_mean	0.0022	0.678
max_y_less_mean	-0.006	0.264

Что можно сделать дальше? Кластеризация!

acc1_100	21.9	22.6	19.3	11.8	19.5
acc2_100	2.9	3.7	2.3	1.6	2.5
acc3_100	0.3	0.6	0.3	0.2	0.3
drg1_100	6.5	8.1	6.7	5.3	7.1
drg2_100	0.9	1.4	0.9	0.8	0.9
drg3_100	0.0	0.0	0.0	0.1	0.0
side1_100	6.7	7.3	6.5	5.1	6.8
side2_100	0.8	1.1	0.8	0.7	0.9
side3_100	0.2	0.3	0.2	0.3	0.2
avg_daily_business_mileage	268.5	290.1	258.6	634.5	182.1
avg_daily_morning_jam_mileage	33.5	33.1	32.3	75.1	23.8
avg_daily_night_mileage	63.7	70.1	61.2	159.8	41.3
avg_speed	24.9	26.5	25.4	36.2	23.7
max_evening_jam_speed	62.2	75.9	63.6	89.0	58.0
max_morning_jam_speed	73.4	79.7	69.1	99.8	63.2
max_night_speed	78.0	86.5	74.0	105.1	67.2
max_speed	117.0	129.3	114.1	139.6	110.0

Оценка влияния ремонтных дорожных работ на частоту и тяжесть ДТП

- Почему это важно и интересно:
  - Аварии – самая частая причина смертей молодых (от 20 до 29 лет) и приводят к потерям до 3% ВВП ежегодно
  - Ремонтные работы меняют характеристики дороги
  - Значит, они могут влиять и на аварии!
  - Необходимо собрать данные по авариям, ремонтам, контрольные переменные – и проверить!

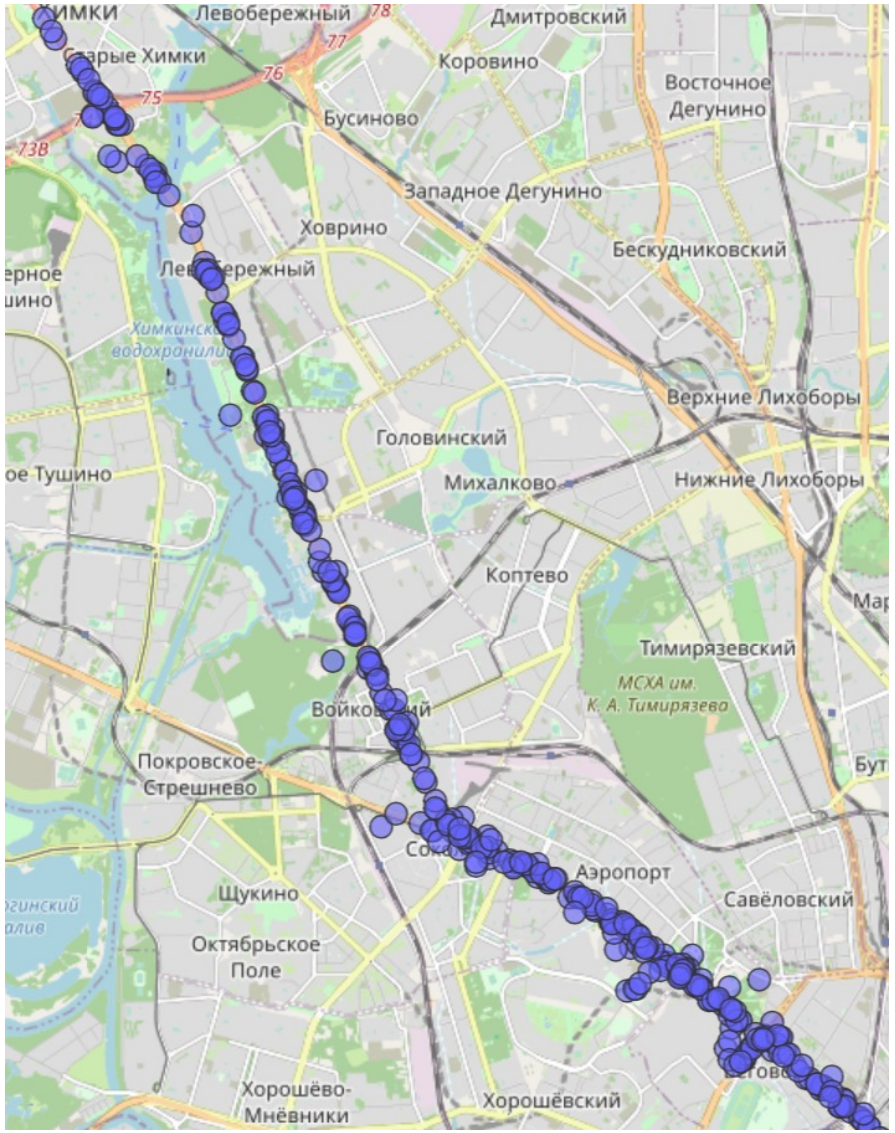
Структура базы данных

- Вручную загружены, соединены и отформатированы 150 XML файлов с карточками ГИБДД.
- С помощью координат отфильтровано более 40000 наблюдений, в результате осталось 500.
- К каждому наблюдению добавлены данные погоды из GSOD.
- Добавлены данные более чем 20 ремонтных работ в этом периоде.

Выборка должна быть однородная!

Начало	Конец
Тверская улица	Метро Белорусская
Мост у метро Белорусская	пересечение с ТТК у метро Динамо
пересечение с ТТК у метро Динамо	Институт Гидропроект у метро Сокол
Институт Гидропроект у метро Сокол	Метро Водный стадион
Метро Водный стадион	Пересечение с МКАД
Пересечение с МКАД	Уровень аэропорта Шереметьево

- Факторы, влияющие на ДТП на маленьких улицах и на больших дорогах отличаются, поэтому было решено исследовать только часть Ленинградского шоссе.
- Наивно предполагать, что факторы ДТП на всей протяжённости шоссе одинаковые.
- Разделение на участки по пересечению с другой крупной транспортной артерией и развязкой увеличивает однородность данных.



Две группы моделей

Логистическая регрессия

- Зависимая переменная: «0» – нет пострадавших, «1» – хотя бы 1 участник получил травмы.
- Распределение по классам: $N_0 = 260$, $N_1 = 213$

Мультиномиальная логистическая регрессия

- Показывает переход от базового класса
- Зависимая переменная: «0» – нет пострадавших, «1» – хотя бы 1 участник находился на амбулаторном лечении или в условиях дневного стационара, «2» – хотя бы один участник находился на стационарном лечении, был тяжело ранен или погиб в результате ДТП.
- Распределение по классам: $N_0 = 260$, $N_1 = 167$, $N_2 = 46$

Панельная Пуассон регрессия для данных о частоте ДТП на участке дороги

- Классический метод анализа счетных данных и частотности ДТП.
- Накладывает ограничение: $\mathrm{Var} = mean$

Панельная отрицательная биномиальная регрессия

- Более современный метод анализа частотности ДТП
- **Не** накладывает ограничение: $\mathrm{Var} = mean$
- Плохо работает с overdispersed данными

Модель для наличия пострадавших в аварии

Переменная	Оценка
Близость к ремонту	0.044
Тип ДТП: наезд на ТС	1.355*
Тип ДТП:Наезд на пешехода	-3.827***
Условия местности: Остановка	1.578*
Объект рядом: Остановка	1.502**
Объект рядом: АЗС	1.643*
Нарушение: перестроение	-2.283***
ROC-AUC	0.95
Уровни значимости: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’	

Модель для тяжести ДТП

	Переход от 0 к 1	Переход от 0 к 2
Близость к ремонту	0.033	-0.056
Наезд на ТС	0.484	2.191**
Наезд на препятствие	119.398***	120.241***
Наезд на пешехода	-3.574***	-52.874***
Падение пассажира	35.716***	36.604***
Опрокидывание	83.375***	86.891***
Дополнительный фактор: мешающее ТС	6.392***	5.709***
Условия местности: подход к мосту	-21.284***	-21.056***
Объект рядом: Остановка	1.320**	1.052
Объект рядом: отделенная парковка	64.293***	66.033***
Мужчина и женщина	89.283***	90.024***
Нарушение: несоблюдение бокового интервала	25.278***	27.197***
Нарушение: перестроение	-1.616**	-32.081***

Уровни значимости: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’

Модель для частоты ДТП на определённом участке

Переменная	Оценка
	Пуассон регрессия
Константа	-
Наличие активного ремонта	0.653***
Среднемесячная температура	0.084*
Кол-во дождливых дней	-
Кол-во дней с моросящим дождем	0.057**
Кол-во дней с покрытием лед	0.403*
Уровни значимости: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’	

Скоринг клиентов компании-арендодателя

- Задача: сделать систему для скоринга клиентов компании
- Данные: слабо структурированные выгрузки данных из ряда сервисов
- Необходимо:
  - Сделать систему показателей на основе имеющихся данных
  - Разработать скоринговую модель

ИСХОДНЫЕ ДАННЫЕ

ФИО: Бор
Дата рождения: 01.10.1982 г.
Паспорт: 4607 167793
Адрес: г Климовск , Симферопольск , кв. 151
Адрес (временный): г Климовск , Симферопольская , д 151

АРХИВНАЯ ИНФОРМАЦИЯ

РФ_ССП_ФЛ+ИП

ФИО	Б
Дата рождения	01.10.1982 г.
Адрес	МОСКОВСКАЯ КЛИМОВСК СИМФЕРОПОЛЬСКАЯ СИМФЕРОПОЛЬСКАЯ КЛИМОВСК МОСКОВСКАЯ КЛИМОВСК СИМФЕРОПОЛЬСКАЯ 45 151 ПОДОЛЬСК КЛИМОВСК МКР СИМФЕРОПОЛЬСКАЯ 151 КЛИМОВСК СИМФЕРОПОЛЬСКОЕ Ш
Паспорт	4607167793
Документ	ОВД Г КЛИМОВСКА УВД Г ПОДОЛЬСКА И ПОДОЛЬ 15.02.2006 15.02.2006 ОТДЕЛОМ ВНУТРЕННИХ ДЕЛ ГОР. КЛИМОВСКА УВД ГОР. ПОДОЛЬСКА И ПОДОЛЬСКОГО РАЙОНА МОСКОВСКОЙ ОБЛАСТИ 502020 4 793 15.02.2006 ОТДЕЛОМ ВНУТРЕННИХ ДЕЛ ГОР. КЛИМОВСКА УВД ГОР.ПОДОЛЬСКА И ПОДОЛЬСКОГО РАЙОНА МОСКОВСКОЙ ОБЛАСТИ 502020 793 15.02.2006 ОВД Г КЛИМОВСКА УВД Г ПОДОЛЬСКА И ПОДОЛЬ
ИНН	

Исполнительное производство

1.	Дата исполнительного производства	12.01.2016
	Орган подразделения ФССП	Подольский РОСП УФССП России по Московской области

1.	Дата исполнительного производства	12.01.2016
	Орган подразделения ФССП	Подольский РОСП УФССП России по Московской области
	№ исполнительного производства	
	Вид исполнительного производства	Штраф ГИБДД
	Долг	500
	Статус	ФЛ
	ОКАТО	45
	Регион	Московская область
2.	Дата исполнительного производства	12.01.2016
	Орган подразделения ФССП	Подольский РОСП УФССП России по Московской области
	№ исполнительного производства	
	Вид исполнительного производства	Штраф ГИБДД
	Долг	500
	Статус	ФЛ
	ОКАТО	45
	Регион	Московская область
3.	Дата исполнительного производства	12.08.2016
	Орган подразделения ФССП	Подольский РОСП УФССП России по Московской области
	№ исполнительного производства	
	Вид исполнительного производства	Штраф ГИБДД
	Долг	500
	Статус	ФЛ

РФ_Межбанковский обмен_2014

ФИО	[redacted]
Дата рождения	[redacted]
Место рождения	КЛИМОВСК МОСК МОСКОВСКОЙ МО
Адрес	МОСКОВСКАЯ КЛИМОВСК СИМФЕРОПОЛЬСКАЯ [redacted]
Паспорт	[redacted]
Документ	[redacted] Отделом внутренних дел гор. Климовска УВД гор. Подольска и Подольского 15.02.2006 [redacted] 15.02.2006
Телефоны	[redacted]
Банки	
ИНФ	КАТЕГОРИЯ: СТОПЛИСТЫ ВТБ24 2014 ТЕЛЕФОН РЕГ.: [redacted] ТЕЛЕФОН ФАКТ. [redacted] ТЕЛЕФОН МОБ.: [redacted] ТЕЛЕФОН РАБ.: [redacted] МЕСТО РАБОТЫ: ООО [redacted] ДОЛЖНОСТЬ: ДИРЕКТОР

ЕГРИП

ФИО	
ИНН	
ОГРНИП	
Тип	Индивидуальный предприниматель
Адрес	МОСКОВСКАЯ 142184, Климовск г Симферопольская у Дата актуальности: 30.06.2015
Телефон	Дата актуальности: 30.06.2015
ОКВЭД основной	Торговля розничная вне магазинов, палаток, рынков
Дата создания	г.
Регистрирующий орган	Межрайонная инспекция федеральной налоговой службы №5 по московской области
Статус	Действующее
Дата актуальности	09.06.2018 г.

Автомобили	
ИНФ	КАТЕГОРИЯ: МОСКВА+ОБЛ_ГАИ_2012 ФИО ВЛАДЕЛЬЦА: ДАТА РОЖДЕНИЯ ВЛАДЕЛЬЦА: МЕСТО РОЖДЕНИЯ ВЛАДЕЛЬЦА: МОСКОВСКАЯ ОБЛАСТЬ ГОР.КЛИМОВСК ПОЛ: МУЖСКОЙ ГРАЖДАНСТВО: РОССИЙСКАЯ ФЕДЕРАЦИЯ ДОКУМЕНТ ВЛАДЕЛЬЦА: Г.КЛИМОВСКА МО АДРЕС ВЛАДЕЛЬЦА ПОЛНЫЙ: КЛИМОВСК ТЕЛЕФОН ВЛАДЕЛЬЦА: РЕГНОМЕР: МАРКА АМТС: БМВ МОДЕЛЬ АМТС: X1 XDRIVE20D ГОД ВЫПУСКА: 2011 VIN: ЦВЕТ: СЕРЫЙ № ДВИГАТЕЛ МОДЕЛЬ ДВИГАТЕЛЯ: N47D20C РЕГ ДОКУМЕНТ: ПТС: № ПОЛИСА ОСАГО: ДАТА ПОЛИСА ОСАГО: ДАТА ПОЛИСА СТРАХОВАЯ КОМПАНИЯ: ОАО "ВСК" МОЩНОСТЬ (ЛС): 177 ОБЪЕМ ДВИГАТЕЛЯ: 1995 СТОИМОСТЬ ТС: 1590000 ТАМОЖЕННЫЕ ОГРАНИЧЕНИЯ: НЕ УСТАНОВЛЕНЫ КОД ОПЕРАЦИИ: 01

Пенсионный фонд	
ИНФ	ГОД ДОХОДА: [illegible] РАБОТОДАТЕЛЬ: МГТУ "СТАНКИН" ИНН РАБОТОДАТЕЛЯ: 7707003506 ТЕЛЕФОН РАБОТОДАТЕЛЯ: 499 9733089 ДОХОД: [illegible] ГОД ДОХОДА: [illegible] РАБОТОДАТЕЛЬ: СТАНКИН ИНН РАБОТОДАТЕЛЯ: 7707003506 ТЕЛЕФОН РАБОТОДАТЕЛЯ: 499 9729455 ДОХОД: [illegible] ГОД ДОХОДА: [illegible] РАБОТОДАТЕЛЬ: МГТУ "СТАНКИН" ИНН РАБОТОДАТЕЛЯ: 7707003506 ТЕЛЕФОН РАБОТОДАТЕЛЯ: 499 9733089 ДОХОД: [illegible] ГОД ДОХОДА. [illegible] РАБОТОДАТЕЛЬ: ГОУ МГТУ "СТАНКИН" ИНН РАБОТОДАТЕЛЯ: 7707003506 ТЕЛЕФОН РАБОТОДАТЕЛЯ: СУММА ГОДОВОГО ДОХОДА: [illegible] ГОД: [illegible] СТРАХОВОЕ СВИДЕТЕЛЬСТВО: [illegible] РАБОТОДАТЕЛЬ: ГОУ ВПО МГТУ "СТАНКИН" ИНН РАБОТОДАТЕЛЯ: 7707003506 ТЕЛЕФОН РАБОТОДАТЕЛЯ: 499 9733840 ДОХОД: [illegible]

Возможные переменные

- Слово «уголовное дело», «уголовное нарушение» найдено в полях информации по клиенту
- В штрафах обнаружен факт управления транспортным средством в нетрезвом состоянии за последние 10 (десять) лет
- Суммарный долг в разделах ФССП (Федеральная служба судебных приставов)
- Является руководителем и/или учредителем более 10 организаций
- Оценка исковой суммы дел в процессе рассмотрения (за 12 месяцев) по отношению к уставному капиталу компании
- *И многие другие…*