## Общий план выступления

Чтобы уложиться во время и не нарушить логику повествования, выступающий будет 1 - (ориентировочно) Арсений. На этапе вопросов, дабы показать целостность команды, выступающий будет редиректить вопросы на других и сам тоже будет отвечать. Темы пересекаются распределять будем ± равномерно

## Ответственные за темы для вопросов

Булдаков А - Дашборд, Преза, выбросы, гипотезы, выводы

Лебедев Ф - Проведение расчётов, гипотезы, заполнение пропусков

Плюхин А - выбросы, гипотезы, дубликаты, выводы, ml, исправления орфографических ошибок в данных

## Выбросы:

- рост - Указан в футах (значения ~ 5.00), Суперизкие женщины из Инди (130-140 см)
- скорость воздуха в помещении - указана в см/с, поэтому выбросы в основном двузначные числа
- температура - сукаФарренгейты

## Заполнение пропусков

(возраст, температура воздуха в помещении и температура воздуха на улице)

- Возраст - медиана по стране
- Температура воздуха в помещении - медина по городу
- Температура воздуха на улице - медиана по городу, году, времени года

## Корреляции

- корреляции год - <что-то> являются особенностью данных (за 2010 собраны данные за один город, за 2011 за другой)
- климат - год, климат - страна: корреляции очевидны (города не перемещаются из одной климатической зоны в другую, по крайней мере в рамках наших данных)
- город - способ_охлаждения: эти данные нам также ничего не дают
- страна - <что-то>: особенности данных
- способ_охлаждение - отопление: это не имеет смысла, поскольку корреляция не видна графическим методом и данные содержат много пропусков. Если эти пропуски убрать, то pvalue значительно поднимется
- ощущение_движения_воздуха_(bool) - ощущение_температуры_(bool): видимо если людям нравится как двигается воздух, то им нравится температура
- ощущение_движения_воздуха_(bool) - предпочтительное_изменение_температуры: если людям нравится как движется воздух, то они не хотят менять температуру?
- занавески - вентилятор: наличие занавесок говорит нам о наличии вентрилятора у человека

## Гипотезы

- **Есть ли разница между средней комфортной температурой для разных возрастных групп** - Да, но разница ± 1 градус, поэтому вывод носит чисто умозрительный характер
- **Влияет ли способ охлаждения на оценку комфорта** - Да, причем клиенты больше довольны вентиляцией, чем кондиционированием
- **Влияние пола на оценку комфорта** - нет
- **Влияет ли возрастная группа на оценку комфорта** - да
- **Взаимосвязь между количеством рекламаций и оценкой комфорта** - Да, взаимосвязь есть, однако она не сильная. Чем больше рекламаций, тем хуже оценка комфорта.
- **средняя оценка комфорта отличается в зависимости от страны** - Да
- **Фактор Среднемесячная температура на улице влияет на оценку комфорта** - Да, но модуль коэфицента корреляции приблизительно равен 0.3, что говорит об очень слабой взаимосвязи
- **Вес респондента влияет на его ощущение температуры** - нет
- **Категории влажности влияют на оценку комфорта** - да
- **Принадлежность респондента к возрастной группе влияет на его ощущение температуры** - да
- **Принадлежность респондента к возрастной группе влияет на его ощущения движения воздуха** - нет
- **есть ли корреляция оценки воздуха с его реальной скоростью?** - да, чем меньше скорость воздуха, тем больше шанс, что респондента устраивает скорость воздуха в помещении

## Выводы

- в сухих климатах (жаркий полузасушливый) люди чаще предпочитают вентиляцию, а во влажных (влажный субтропический муссонный, тропическая влажная саванна) - кондиционирование
- Видно как клиенты не довольны продукцией в Техасе, стоит обратить внимание на этот рынок.
- Способ охлаждения влияет на скорость воздуха, как было доказано выше, чем меньше скорость воздуха, тем больше клиентов довольны. Возможно стоит обратить внимание на то, как работает кондиционер, чтобы уменьшить скорость воздуха, или устанавливать его в места, где нет людей, чтобы на них "не дуло". Это могло бы быть хорошим местом для инвестиций, так как получится закрыть потребности клиентов новым инженерным решением.

# Текст

1. Слайд 1
    a. Добрый день, уважаемые слушатели
    b. Наша тема - анализ данных в сфере бытовых услуг
2. Слайд 2
    a. На этапе обработки данных мы столкнулись с аномалиям в данных. Это, во-первых, опечатки, во-вторых, выбросы. БОльшую часть выбросов нам удалось объяснить и восстановить. Например, рост указанный в футах или температура в Фаренгейтах.
    b. Также мы заполнили пропуски в возрасте, скорости воздуха, температуре в помещении и на улице, потому что это числовые столбцы и в них довольно мало пропусков.
3. Слайд 3
    a. После проведения расчётов и выдвижения гипотез мы решили проанализировать как отличается средняя оценка комфорта респондентов в зависимости от их климата и способа охлаждения. Для это мы ввели новый категориальный столбец и построили ящик с усами. На нём мы заметили некоторые особенности, но для большей ясности решили перестроить график
4. Слайд 4
    a. Вот здесь уже отчетливей видно, что в сухих климатах Кондиционирование получает оценку меньше чем вентиляция. Чтоб это было наглядней видно, мы построили сводную табличку и вот тут разница становится очевидней. Но почему же так происходит?
5. Слайд 5
    a. Здесь в дело вступает физика, потому что в более холодном воздухе, давление насыщенного пара ниже. При понижении температуры в воздухе образуется избыток влаги, и этот избыток конденсируется на радиаторе кондиционера. Поэтому кондиционер сушит воздух, что естественно не нравится людям, живущим в засушливых климатах, потому что там у них воздух и так сухой.
    b. (Давление насыщенного пара это давление пара, при котором жидкость находится в равновесии со своим паром. $P=nkT$. где n - концентрация молекул пара k - постоянная Больцмана T - температура.)
6. Слайд 6
    a. Далее мы задумались: чем еще может отличаться кондиционирование от вентиляции. Мы начали рассматривать различные факторы в нашем датасете и заметили, что кондиционер делает скорость воздуха больше, чем вентиляция, эту корреляцию прекрасно видно на слайде. Возможно скорость воздуха как-то влияет на конечную оценку? Ведь никому не нравится работать при сквозняке кондиционера.

        Так и оказалось: люди предпочитают низкую скорость воздуха в помещении, это можно увидеть по зависимости факторов **скорость воздуха** с **ощущением движения воздуха** и **скорость воздуха** с **оценкой комфорта**, графики зависимостей вы можете увидеть на слайде.

Обобщая сказанное ранее: *если бы системы кондиционирования и вентиляции имели бы возможность работать при более низкой скорости воздуха в помещении, то клиенты были бы более довольны работой нашей компании.*

Поэтому мы считаем, что это может быть подходящем местом для инвестиций, вероятно возможно найти более удачное инженерное решение, которое поможет снизить скорость воздуха в помещении и, соответственно, повысить оценку комфорта наших клиентов. Как пример: если бы хладагент кондиционера мог бы поглощать больше тепла помещения за меньшее время, то мы могли бы уменьшить скорость вентилятора, который выдувает холодный воздух из кондиционера, при этом температура помещения все равно бы уменьшалась за счет более эффективного хладагента, то есть более эффективного обмена тепла.

7. Слайд 7
    a. Выполняя пункты технического задания, мы конечно дошли и до регрессионной модели. Первое, что мы сделали, это посмотрели еще раз на имеющиеся корреляции, которые мы установили в прошлых пунктах задания. Нашей задачей было найти те факторы, которые имеют сильную связь со скоростью воздуха в помещении. Таким образом нами были выбраны следующие факторы: время года, город, способ охлаждения и среднемесячная температура на улице, на слайде вы можете увидеть корреляцию с некоторыми факторами. Даже интуитивно понятно, что эти факторы влияют на температуру воздуха в помещении. Также они не являются труднодоступными, то есть в теории, наша модель могла бы предугадывать температуру в помещении используя такие входные данные, которые совершенно не трудно найти в интернете.
8. Слайд 8
    a. Что же мы получили в итоге? На слайде вы можете увидеть график, который показывает, как обученная модель отработала на случайной выборке нашего датасета: красная линия показывает идеальную модель, то есть 100% попадание, наша модель никогда не промахивается и предсказывает температуру точно. Синие точки – это предсказания нашей обученной модели по факту, синяя линяя показывает корреляцию между предсказанными значениями и тестовыми значениями. Как вы видите, наша модель получилась довольно неточной и часто ошибается хоть и не очень значительно: MAE или среднее абсолютная ошибка равна 1.06, то есть в среднем наша модель ошибается на +-1 градус. Метрики *MSE* и *RMSE* также отражают точность модели, это *среднеквадратичную разница между прогнозируемыми и фактическими значениями, и корень из MSE.* Ориентируясь на эти метрики можно понять насколько наша модель точна.

    ***Если спросят подробности***:
    https://www.codecamp.ru/blog/mse-vs-rmse/

R в квадрате это еще одна метрика, с помощью которой можно понять точность готовой модели: 1 – это идеальная модель, 0 – модель плохая. Резюмируя: у нас получилась не очень точная модель, нам кажется, что это вызвано довольно ограниченным количеством данных в нашем датасете.

9. Слайд 9
    a. Вот собственно наш ДашБорд. Здесь мы вывели основные показатели, связанные с оценкой комфорта: Средняя оценка по возрастной группе, средняя оценка по стране, соотношение климатов и нижняя диаграмма - это средняя оценка от климата и способа охлаждения. Также мы добавили среднее количество рекламаций по стране.
10. Слайд 10
    a. Так теперь к выводам
    b. Первое. В более засушливых климатах следует инвестировать в вентиляцию. В перспективе, конечно, следует вложиться в разработку таких моделей кондиционеров, которые будут сушить воздух меньше или не сушить вовсе.
    c. Второе. Следует инвестировать в разработку систем охлаждения с более низкой скоростью воздуха
    d. Ну и третье. Следует решить вопрос с Техасом. Там самая низкая средняя оценка комфорта. Возможно стоит обратить внимание на модели вентиляций и кондиционеров, которые там используются, возможно стоит обратить внимание на их техническое состояние. Но это требует дополнительных данных и исследований.
11. Слайд 11
    a. Спасибо за внимание.