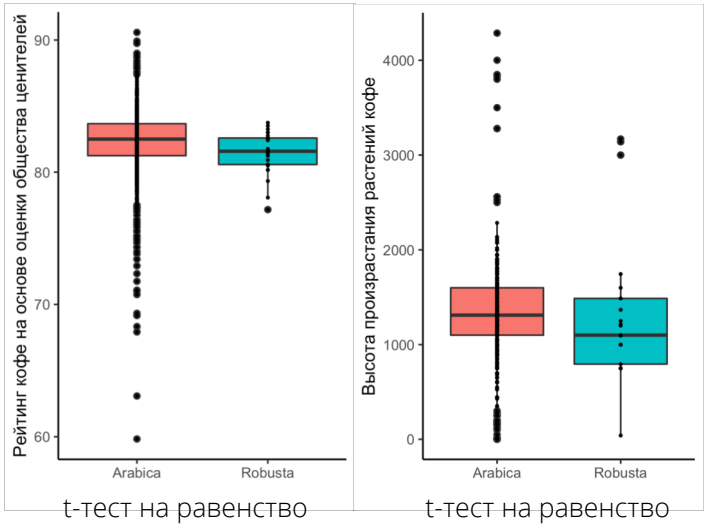
# Анализ датасета Coffee ratings

Павел Никитин, 3 курс ФББ МГУ

Датасет представляет из себя большое количество информации о разных сортах, видах, производителях кофе и его вкусовых характеристик. Для анализа были взяты только самые важные, на мой взгляд показатели. Из датасета были удалены строки, в которых отсутствовала информация по этим показателям.

Для приготовления кофе используются два вида растений кофе: аравийский (*C. arabica*) и конголезский (*C. robusta*).

## Гипотеза 1: У арабики и робусты не отличаются средние значения рейтинга и высоты произрастания растений.

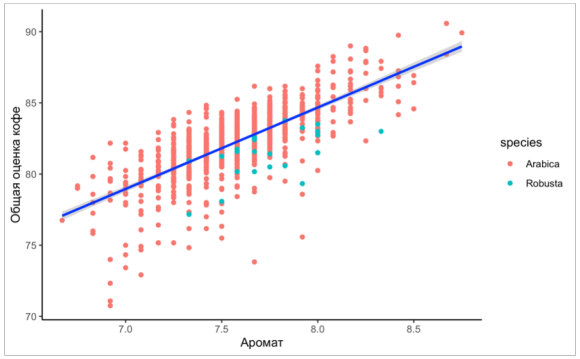



t-тест на равенство средних:
p-value = 0.03699

t-тест на равенство средних:
p-value = 0.763

Вывод: У арабики и робусты значимо не отличается средняя высота произрастания, но отличается средняя рейтинговая оценка.

## Гипотеза 2: Общая оценка кофе связана с его ароматом



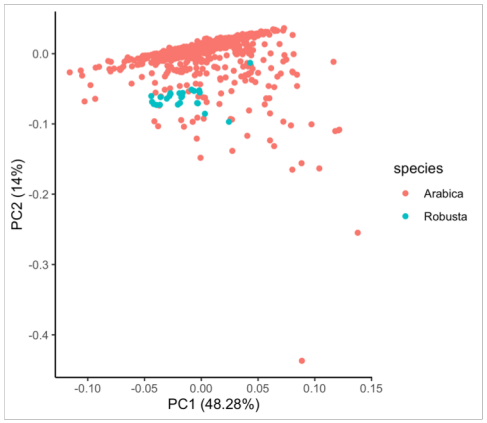

Построим линейную регрессию зависимости оценки от аромата

ax+b
p-value (a) $<2*10^{-16}$
p-value (b) = 0.21
$R^2$ = 0.5455
p-value: $< 2.2*10^{-16}$

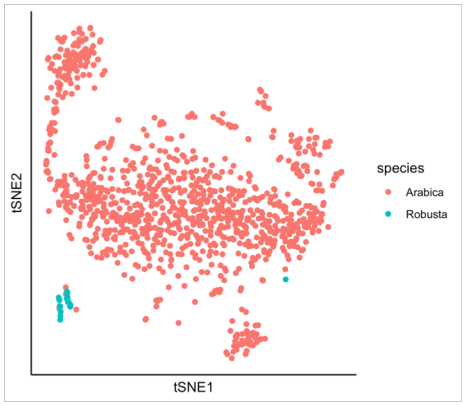
Вывод: Оценка кофе имеет значимую корреляцию с его ароматом.

## Гипотеза 3: Данные о многих параметрах кофе (вкус, аромат, интенсивность следов на кружке, ...) можно кластеризовать на арабику и робусту.

PCA




tSNE




Вывод: Видимо, различия в параметрах напитка обусловлены видом кофе (арабика/робуста), хотя есть исключения.