**Отчет по практикуму (7 – 12)**

Для выполнения заданий на Python я использовал **Jupyter Notebook**. Для выполнения заданий на R использовал **RStudio**.

**7. Продемонстрировать применение для проверки различных гипотез и различных доверительных уровней (0.9, 0.95, 0.99) следующих критериев:**

**a) Стьюдента, включая односторонние варианты, когда**

**проверяемая нулевая гипотеза заключается в том, что одно из сравниваемых средних значений больше (или меньше) другого. Реализовать оценку мощности критериев при заданном объеме выборки или определения объема выборки для достижения заданной мощности;**

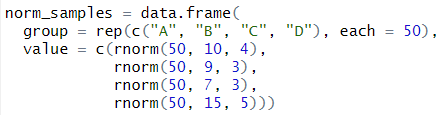
**b) Уилкоксона-Манна-Уитни (ранговые);**

**c) Фишера, Левене, Бартлетта, Флигнера-Килина (проверка**

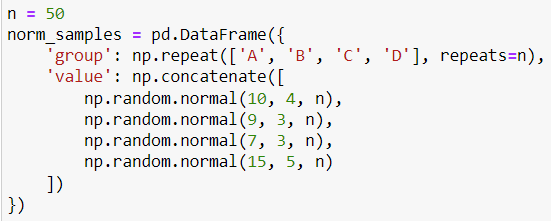
**гипотез об однородности дисперсий).**

**a) Односторонний вариант критерия Стьюдента** служит для проверки нулевой гипотезы о **равенстве** среднего значения выборки некоторому известному значению. Для применения критерия необходимо, чтобы данные имели нормальное распределение. Сгенерируем данные:

**R:**

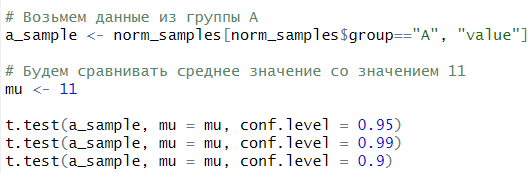
****

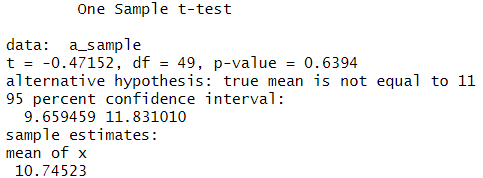
**Python**:

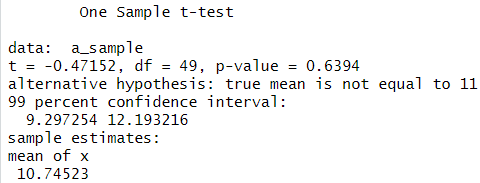
****

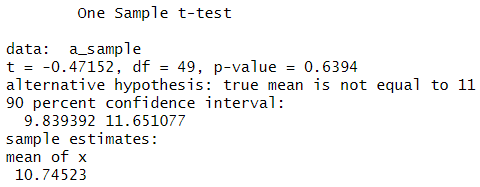
**R:**

Возьмем данные из группы А, у которой матожидание равно 10

****



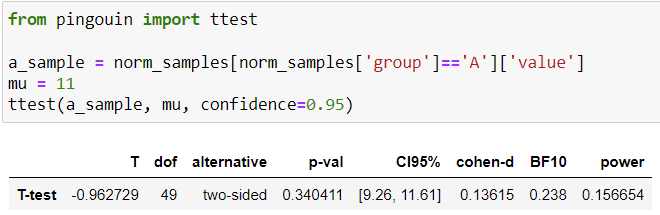


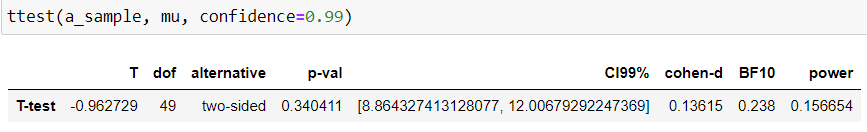


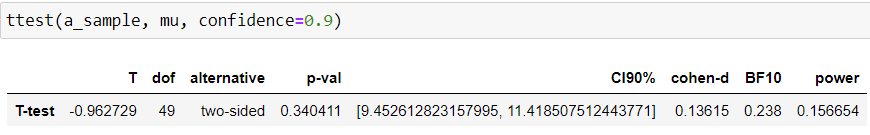
Видим, что *p-value = 0.6394 > 0.05*. Значит, мы не можем отклонить

нулевую гипотезу и заключить, что среднее значение не равно 11.

**Python:**

****

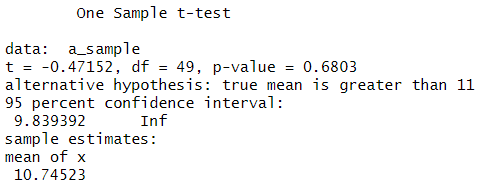
****

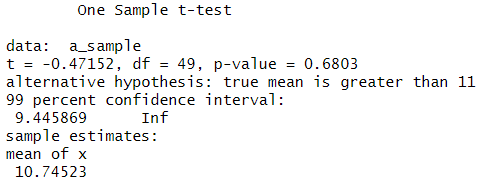
****

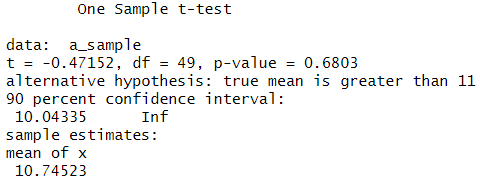
Теперь проверим нулевую гипотезу о том, что среднее значение **больше** одиннадцати.

**R:**



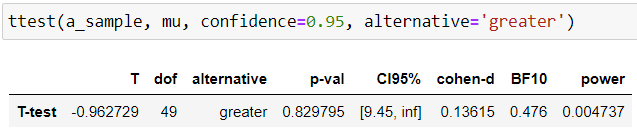




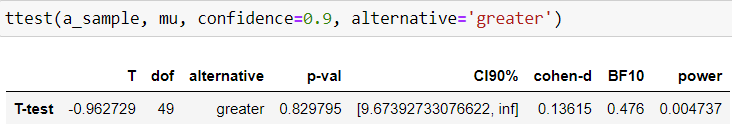


В этом случае тоже нельзя отвергнуть нулевую гипотезу. Мы не можем утверждать, что среднее значение больше одиннадцати.

**Python:**



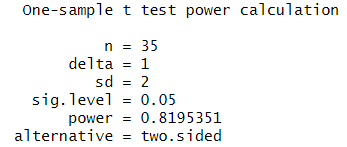




Теперь реализуем **оценку мощности критериев** при объеме выборки, равным 35. Пусть ожидаемая разница в средних значениях равна 1, стандартное отклонение = 2 и уровень значимости = 0.05.

**R:**

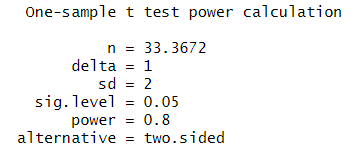




Видим, что мощность критерия оценена на 0.8195351.

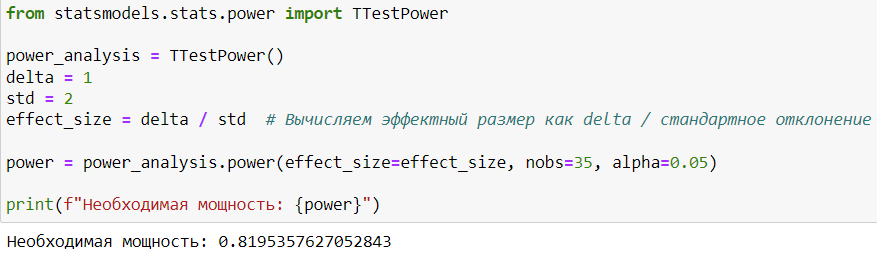
Теперь **определим объем выборки для достижения мощности**, равной 0.8.

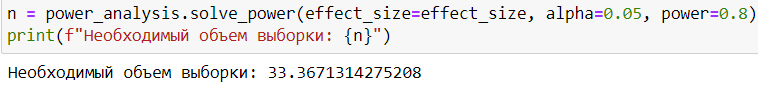




Видим, что объем выборки должен быть больше 33.

**Python:**

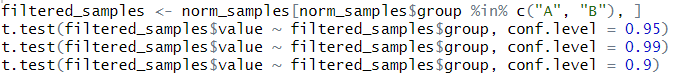


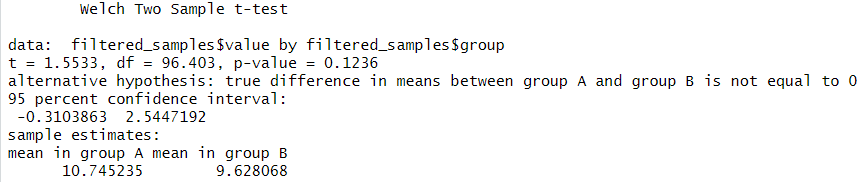


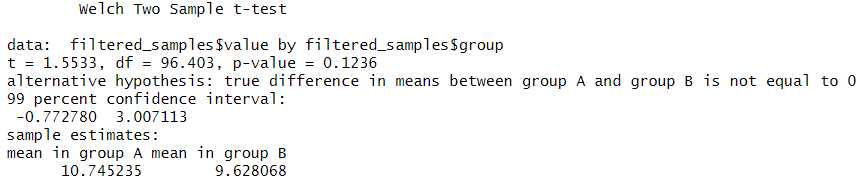
Теперь проведем **тест Cтьюдента для двух независимых выборок**. Нулевая гипотеза состоит в том, что средние значения этих выборок совпадают.

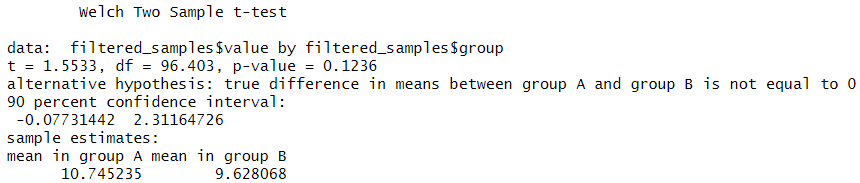
В качестве второй выборки возьмем данные из группы В, у которой матожидание равно 9.

**R:**



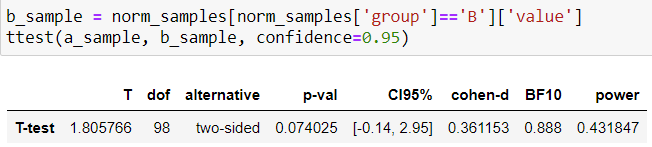


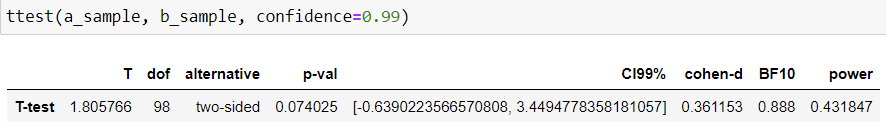


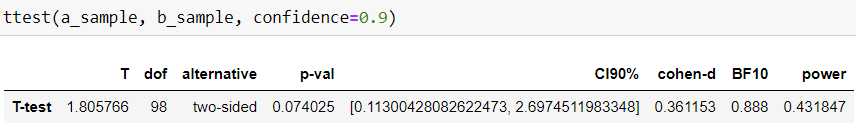


Здесь *p-value = 0.1236 > 0.05.* Значит, мы не можем утверждать, что средние значения отличаются.

**Python:**

****

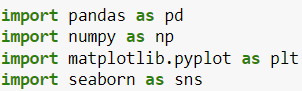
****

****

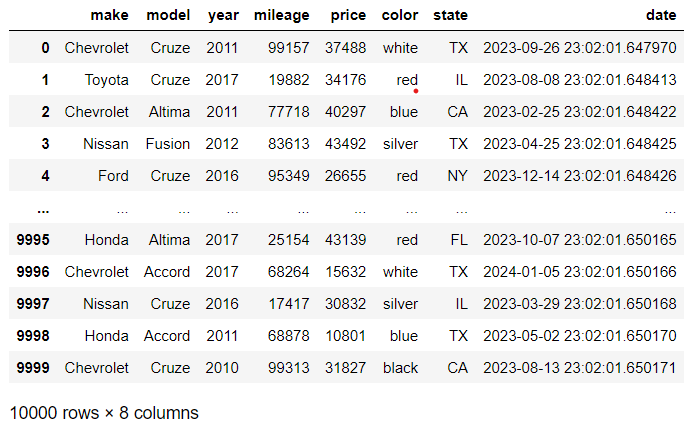
**b)** В случаях, когда данные распределены не нормально, вместо критерия Стьюдента следует использовать его непараметрический аналог - **критерий Уилкоксона-Манна-Уитни.**

В качестве датасета, как и для заданий 1-6, я взял цены автомобилей Chevrolet, Toyota, Nissan, Ford и Honda в США. Помимо цены, в таблице содержится информация о годе выпуска, пробеге, цвете машины, а также о дате продаже, и в каком штате происходила сделка. Вот ссылка на данные:

<https://www.kaggle.com/datasets/at3191/us-car-prices>

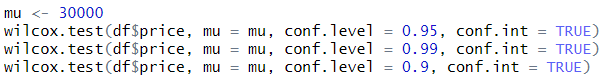


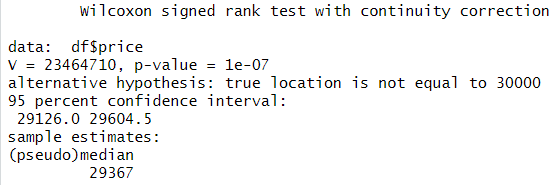


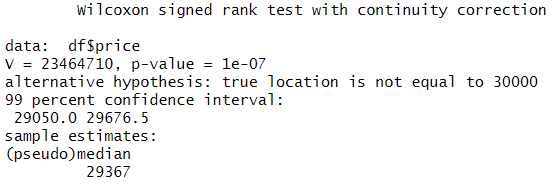


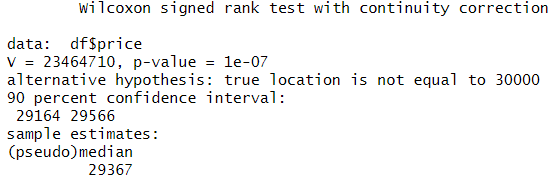
**В одностороннем варианте** сравним среднее значение цен автомобилей со значением, равным 30 000.

**R:**



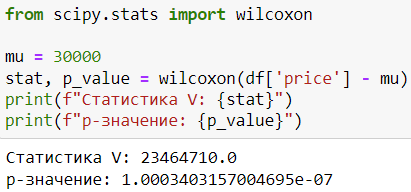






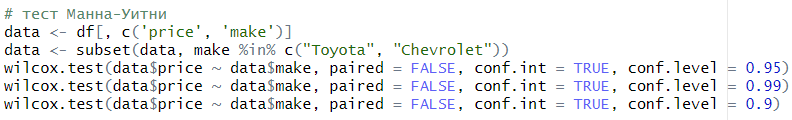
Мы отвергаем нулевую гипотезу и утверждаем, что среднее значение цен автомобилей не равно 30 000.

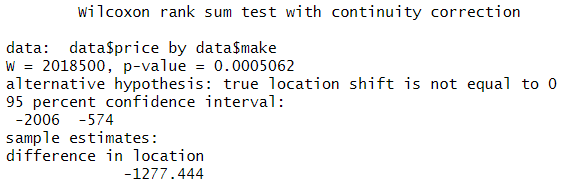
**Python:**

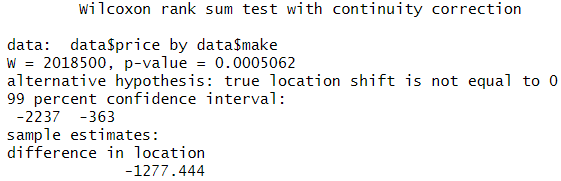


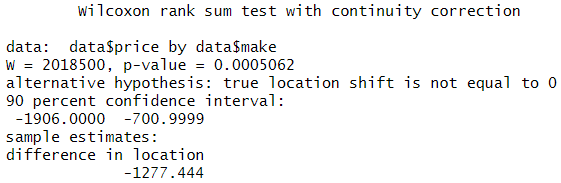
Теперь с помощью **двустороннего варианта** сравним цены автомобилей Toyota и Chevrolet. Нулевая гипотеза состоит в том, что центры распределений, из которых происходят сравниваемые выборки, смещены относительно друг друга на некоторую величину.

**R:**



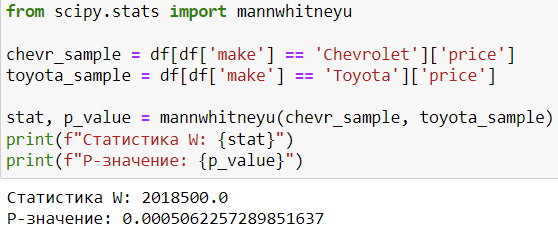






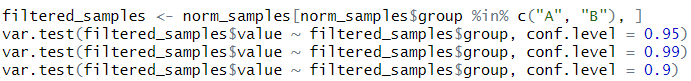
В нашем случае *p-value = 0.00005062 <<0.05.* Cледовательно, мы отвергаем нулевую гипотезу и делаем вывод, что цены автомобилей Toyota и Chevrolet статистически различаются.

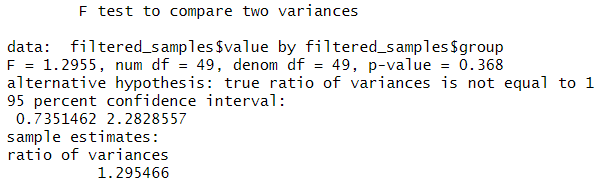
**Python:**

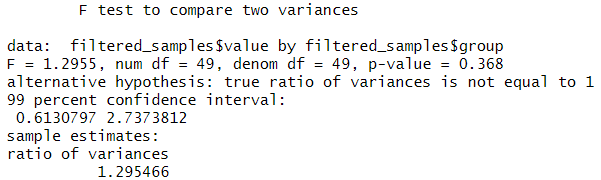


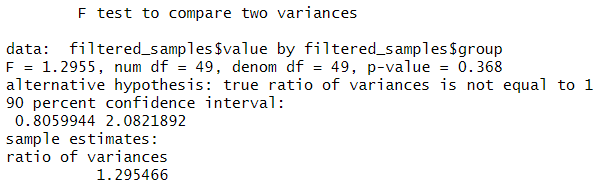
**c) Критерий Фишера** используется для сравнения дисперсий двух нормально распределенных выборок, то есть проверятся нулевая гипотеза о равенстве дисперсий. Сравним дисперсии данных из групп А и В.

**R:**







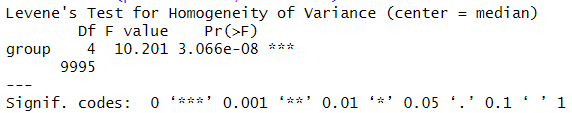


Здесь мы не отклоняем нулевую гипотезу. Мы не можем утверждать, что дисперсии двух групп статистически различаются.

C помощью **критерия Левене** выясним, отличаются ли дисперсии цен автомобилей разных производителей.

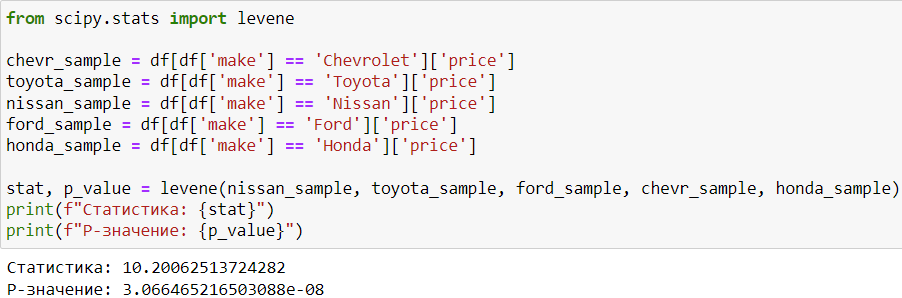
**R:**





В данном случае *Pr(>F) = 3.066e-8 << 0.05*. Следовательно, мы отклоняем нулевую гипотезу и утверждаем, что дисперсии цен автомобилей разных производителей статистически различаются.

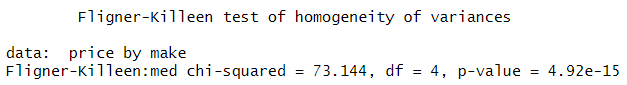
**Python:**

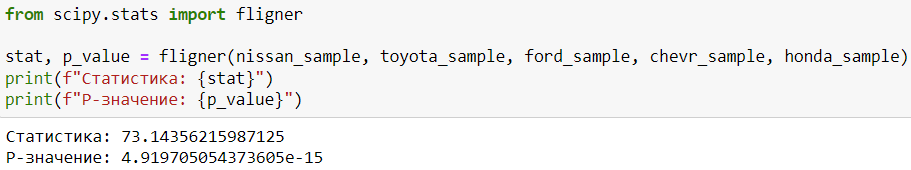
****

Теперь проверим это с помощью **критерия Флигнера-Килина**.

**R:**





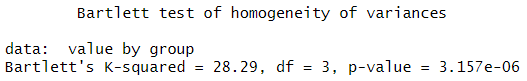
**Python:**

В результате теста можно заключить, что дисперсии цен автомобилей разных производителей статистически различаются.

**Критерий Бартлетта** чувствителен к отклонениям от нормы. Проверим с помощью него, равны ли дисперсии групп А, B, C, D из нормально распределенных данных.

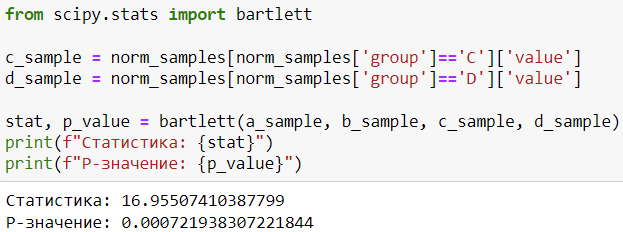
**R:**





Видим, что *p-value << 0.05.* Значит, мы отвергаем нулевую гипотезу и заключаем, что дисперсии групп статистически различаются.

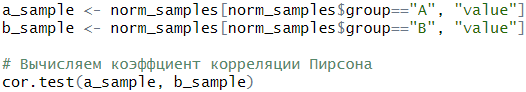
**Python:**

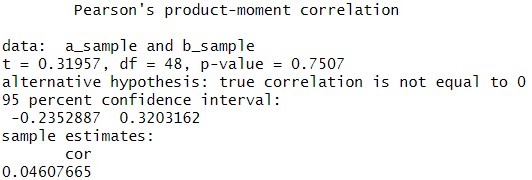
****

**8. Исследовать корреляционные взаимосвязи в данных с помощью коэффициентов корреляции Пирсона, Спирмена и Кендалла.**

Для того, чтобы вычислить **коэффициент корреляции Пирсона** необходимо, чтобы обе анализируемые переменные были распределены нормально. Для этого сгенерируем выборку из нормального распределения.

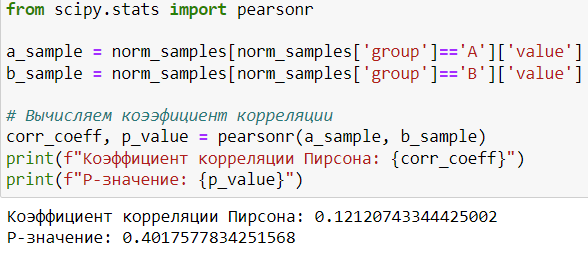
**R:**





Коэффициент корреляции Пирсона получился равен 0.04607665. Тест вычисляет оценку статистической значимости коэффициента, проверяя нулевую гипотезу о равенстве его нулю. В нашем случае *p-value = 0.7507 > 0.05 (уровень значимости)***.** Следовательно, мы не отвергаем нулевую гипотезу, и заключаем, что у нас нет оснований считать, что коэффициент корреляции получился статистически значимым.

**Python:**

****

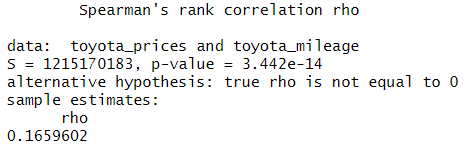
Для ненормально распределенных переменных, а также при наличии нелинейной связи между переменными, следует использовать непараметрический **коэффициент корреляции Спирмена**.

В тесте нулевая гипотеза состоит в том, что отношение шансов равно 1

Выясним, насколько сильно связаны цена автомобилей Toyota с пробегом.

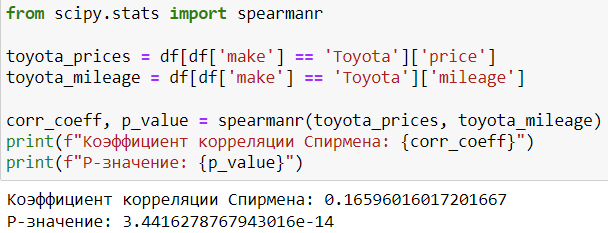
**R:**





*p-value << 0.05*, следовательно, отвергаем нулевую гипотезу и заключаем, что результат получился статистически значимым, т.е между ценой автомобиля Toyota и пробегом есть статистическая связь.

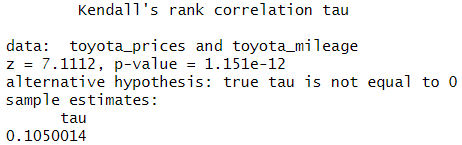
**Python:**

****

Теперь вычислим **коэффициент корреляции Кендалла.**

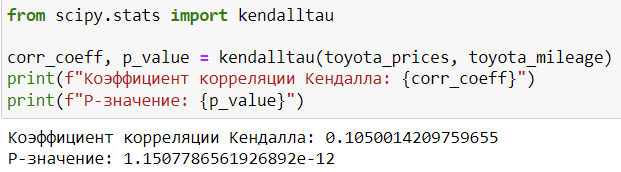
**R:**

****

****

Видим, что результат получился статистически значимым.

**Python:**

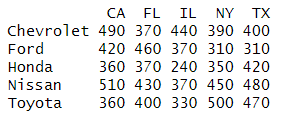


**9. Продемонстрировать использование методов хи-квадрат, точного теста Фишера, теста МакНемара, Кохрана-Мантеля-Хензеля.**

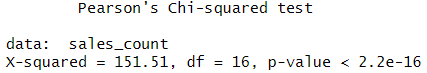
C помощью критерий **хи-квадрат** определим, есть ли статистически значимая взаимосвязь между штатом продажи и маркой автомобиля, который был продан. Иными словами, влияет ли штат на количество продаж той или иной марки автомобиля.

**R:**



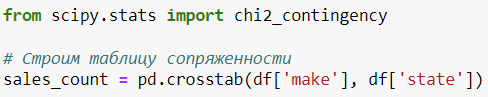


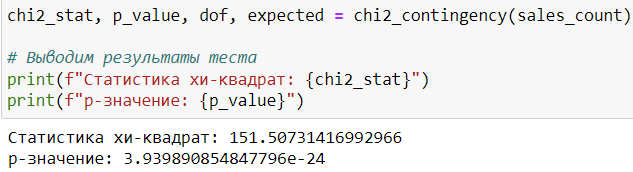




Результат получился статистически значимым, cледовательно, мы можем отклонить нулевую гипотезу об отсутствии влияния штата продажи.

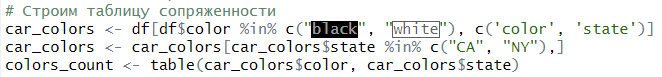
**Python:**

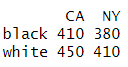
****

****

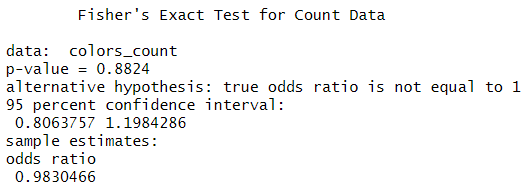
Теперь с помощью **точного теста Фишера** проверим, есть ли между штатом продажи (Калифорния, Нью-Йорк) и цветом автомобиля (черный, белый) статистическая связь.

**R:**



****

****

****

В данном случае *p-value = 0.8824 > 0.05*. Значит, мы не можем отклонить нулевую гипотезу, и у нас нет оснований заключать, что между штатом продажи и цветом автомобиля есть статистическая связь.

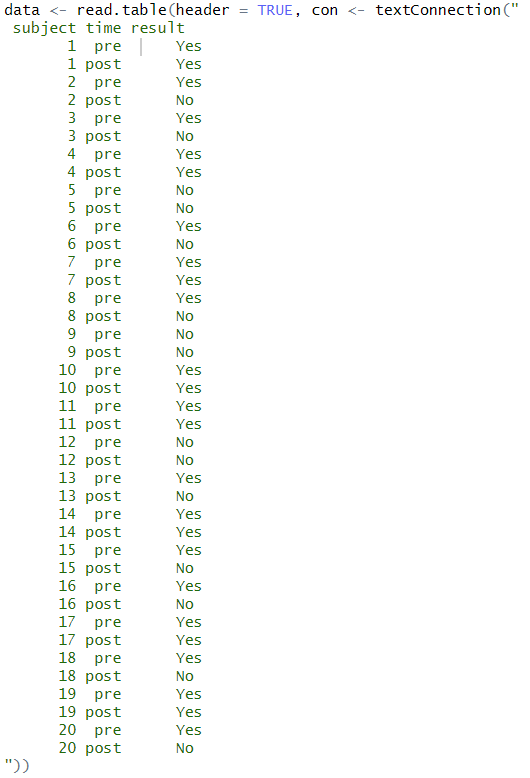
**Python:**

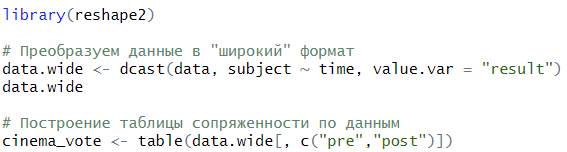
****

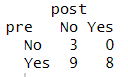
Теперь смоделируем ситуацию: представим, что за месяц до премьеры фильма «Трансформеры: Восхождение Звероботов» случайным образом опросили 20 человек, пойдут ли они в кинотеатр на этот фильм. После выхода фильма появились отзывы критиков, которые оказались преимущественно негативными.

Снова этих же 20 персон было опрошено, пойдут ли они в кинотеатр. С помощью теста **МакНемара** выясним, повлияли ли отзывы критиков на выбор людей.

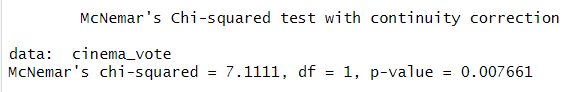
**R:**



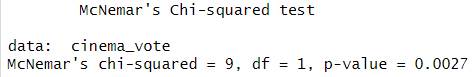






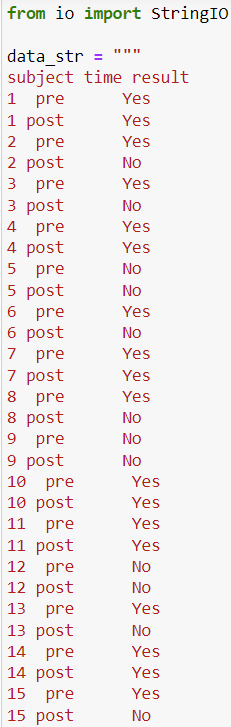


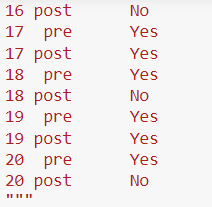


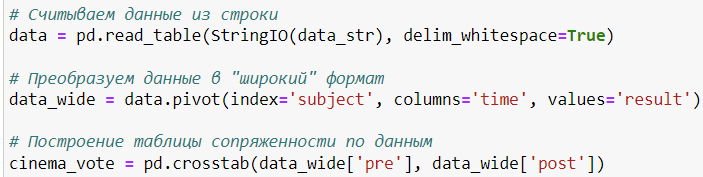


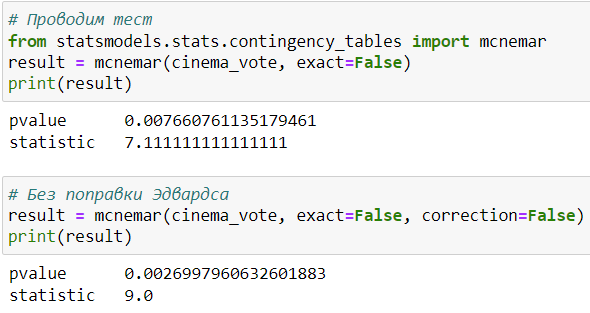
Видим что по итогам обоих тестов результат получился статистически значимым. Поэтому мы можем заключить, что критики повлияли на выбор людей, пойти в кино или нет.

**Python:**



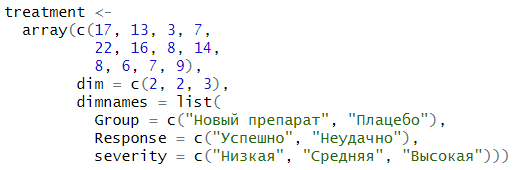


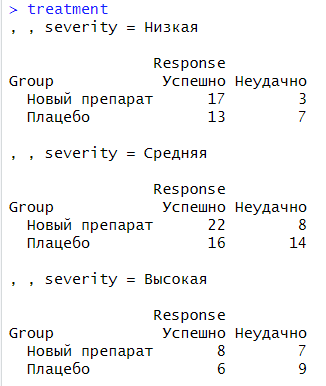




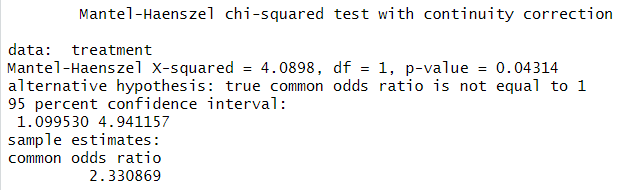
Теперь представим следующее: две группы больных подвергались лечению при помощи нового препарата и плацебо. Испытания проводятся при разных степенях тяжести (низкая, средняя и высокая). По разным причинам можно ожидать, что результаты эксперимента будут несколько отличаться в зависимости от степени тяжести. В конце эксперимента учтено количество случаев успешного и неуспешного лечения в обоих группах. С помощью теста **Кохрана-Мантеля-Хензеля** выясним, является ли исход лечения статистически значимым в зависимости от того, получали ли они новый препарат или нет.

**R:**

****



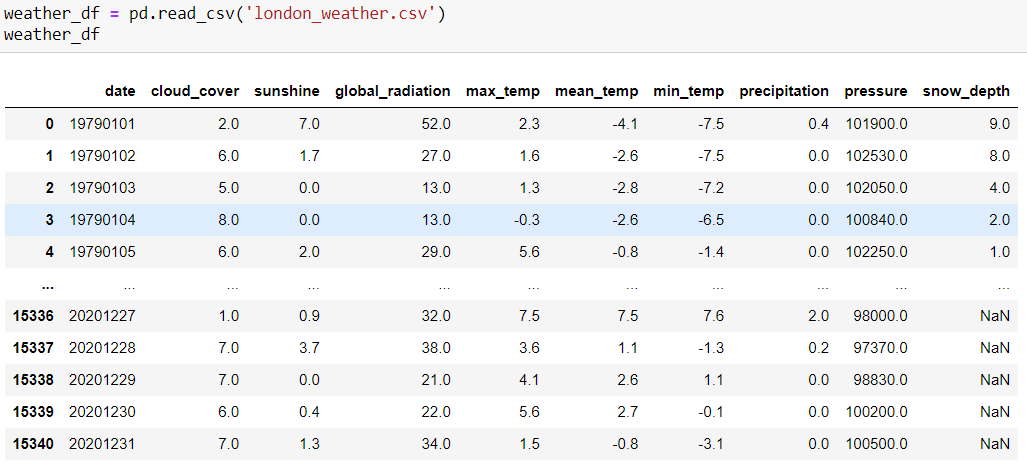




Как видим, исход лечения оказался статистически значимо связанным с тем, получали ли испытуемые новый препарат (p-value = 0.04314).

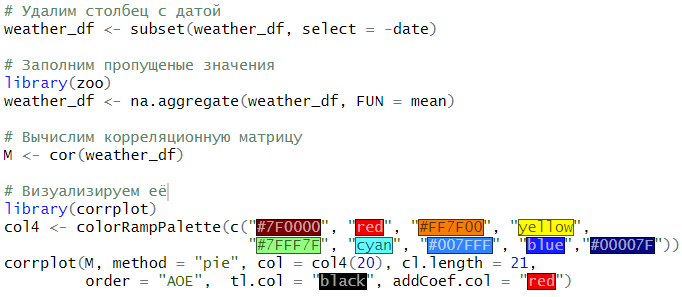
**10. Проверить наличие мультиколлинеарности в данных с помощью корреляционной матрицы и фактора инфляции дисперсии.**

Для выполнения этого задания возьмем датасет с информацией о погоде в Лондоне с 1979 по 2020 годы. Он содержит в себе различные данные: облачность, количество солнечных часов, глобальная радиация, температура, количество осадков, давление и глубина снега. Ссылка на данные: <https://www.kaggle.com/datasets/emmanuelfwerr/london-weather-data>

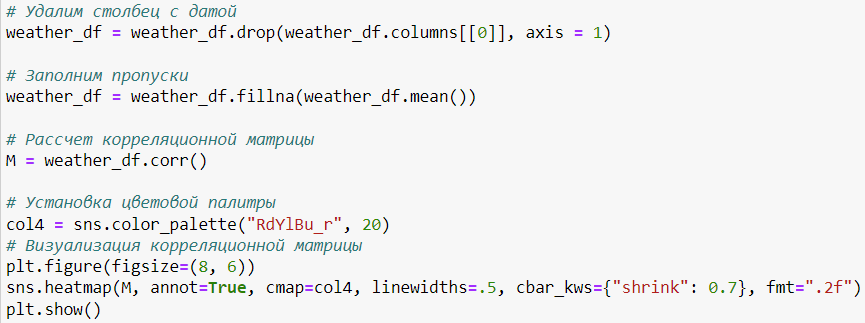


Для того, чтобы вычислить **корреляционную матрицу**, удалим столбец с датой и заполним пропуски средним значением.

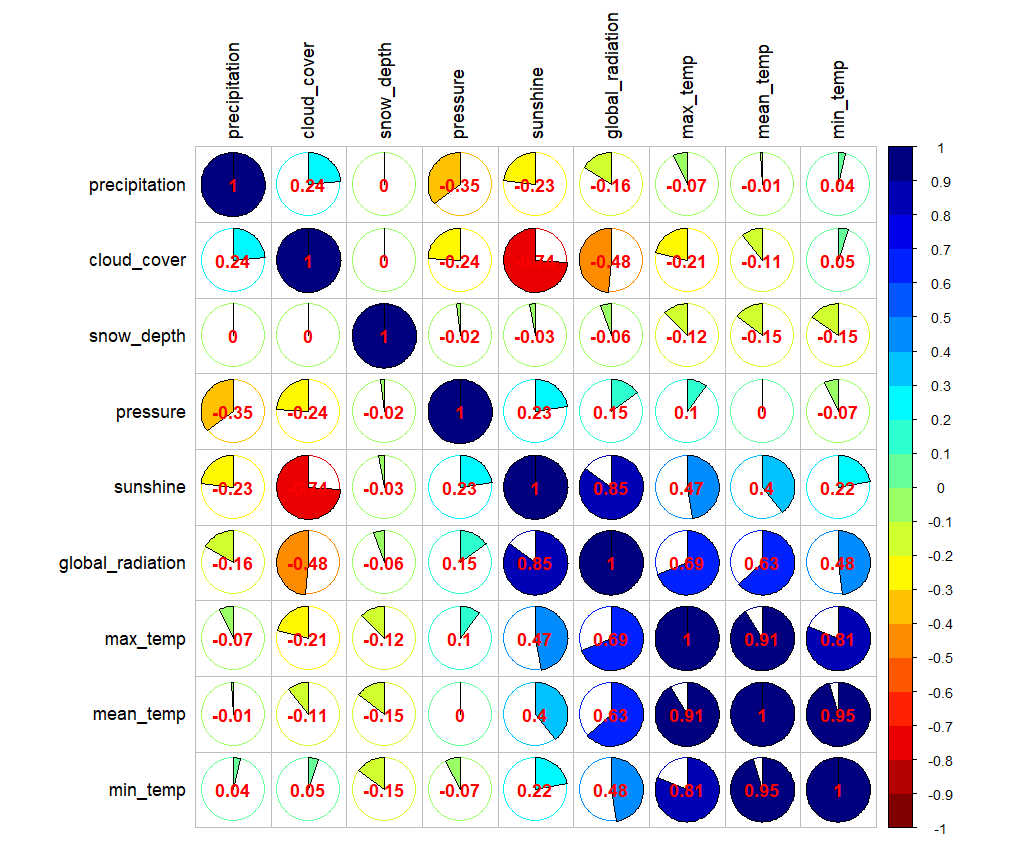
**R:**



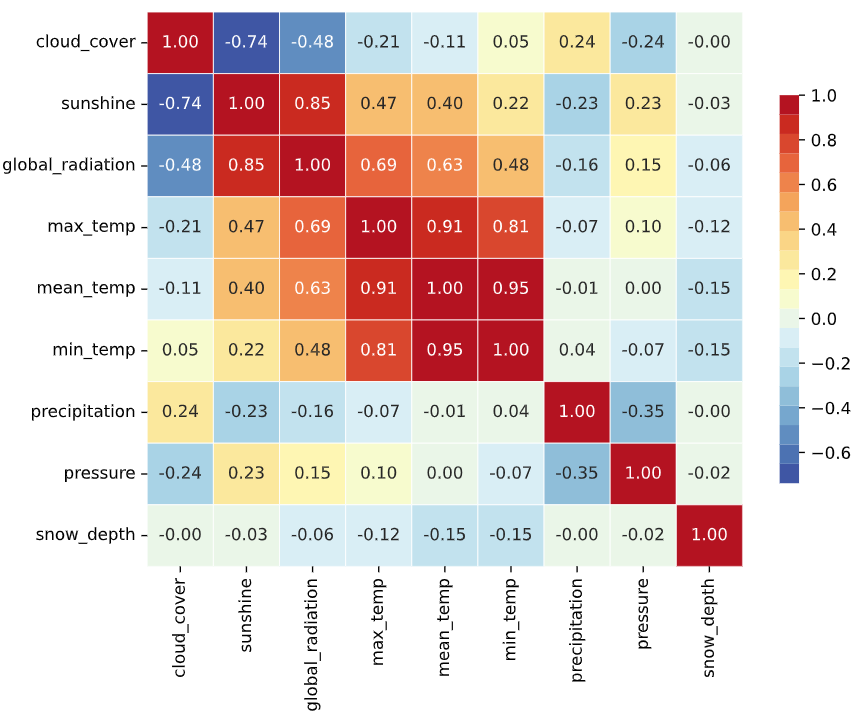
**Python:**



**Корреляционная матрица (R)**



**Корреляционная матрица (Python)**

****

**Видим, что есть признаки с большими коэффициентами корреляции.**

Теперь построим линейную регрессионную модель, где **global\_radiation** предсказывается на основе **cloud\_cover, sunshine, max\_temp, mean\_temp и min\_temp.**

Затем вычислим множественные коэффициенты инфляции дисперсии (VIF) для каждой из переменных-предикторов в этой модели.

**Следующую информацию я узнал по ссылке:** [**https://www.codecamp.ru/blog/variance-inflation-factor-r/**](https://www.codecamp.ru/blog/variance-inflation-factor-r/)

*Значение VIF начинается с 1 и не имеет верхнего предела. Общее эмпирическое правило для интерпретации VIF выглядит следующим образом:*

*Значение 1 указывает на отсутствие корреляции между данной переменной-предиктором и любыми другими переменными-предикторами в модели.*

*Значение от 1 до 5 указывает на умеренную корреляцию между данной переменной-предиктором и другими переменными-предикторами в модели, но часто она недостаточно серьезная, чтобы требовать внимания.*

*Значение больше 5 указывает на потенциально сильную корреляцию между данной переменной-предиктором и другими переменными-предикторами в модели. В этом случае оценки коэффициентов и p-значения в выходных данных регрессии, вероятно, ненадежны.*

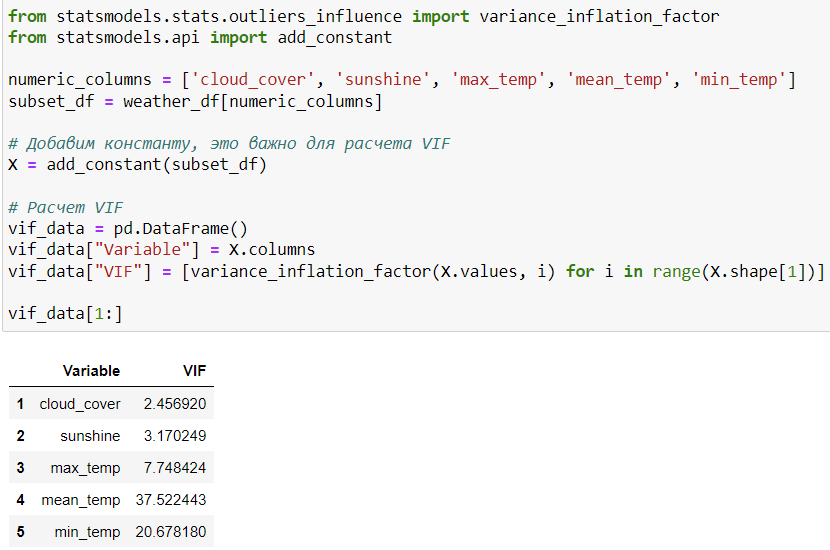
**R:**

****

****

Видим, что мы имеем дело с **выраженной мультиколлинеарностью**, так как значение VIF (**mean\_temp)** = 37.52 > 5, что логично, ведь средняя температура сильно зависит от значений максимальной и минимальной температур.

**Python:**

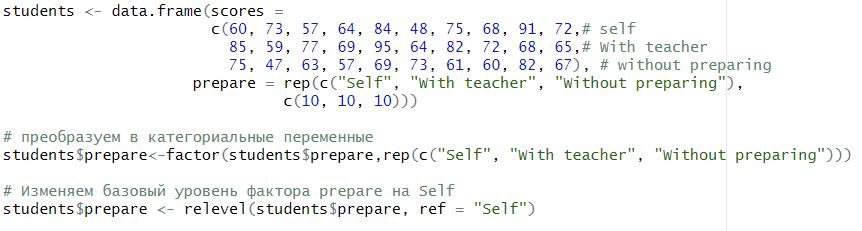
****

**11. Исследовать зависимости в данных с помощью дисперсионного анализа.**

**Однофакторный дисперсионный анализ** используется для определения того, существует ли статистически значимое различие между средними значениями трех или более независимых групп.

Пусть у нас есть 3 группы учеников, которые писали одну и ту же контрольную работу, но способ их подготовки был разным. Первая группа учеников готовилась самостоятельно (без учителя), вторая – с учителем, а третья вовсе не готовилась.

**R:**

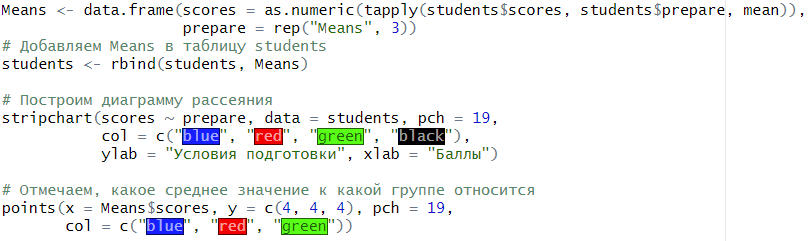
****

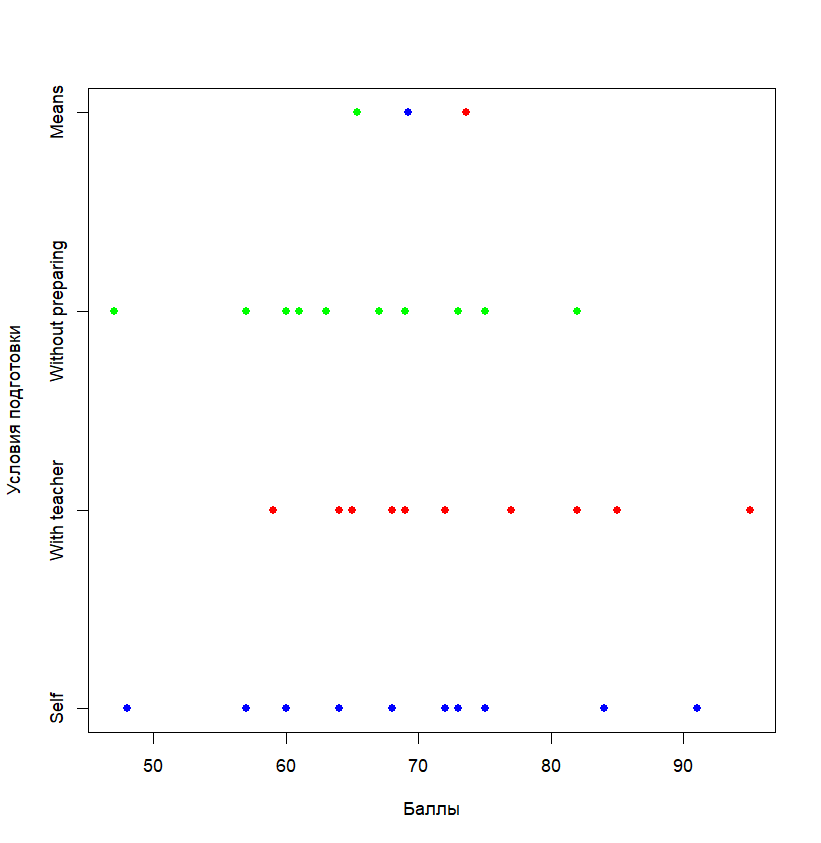
Подлежащую проверке нулевую гипотезу можно сформулировать так: способ подготовки учеников не оказывают никакого влияния на полученные баллы на контрольной работе.





Построим диаграмму рассеяний:



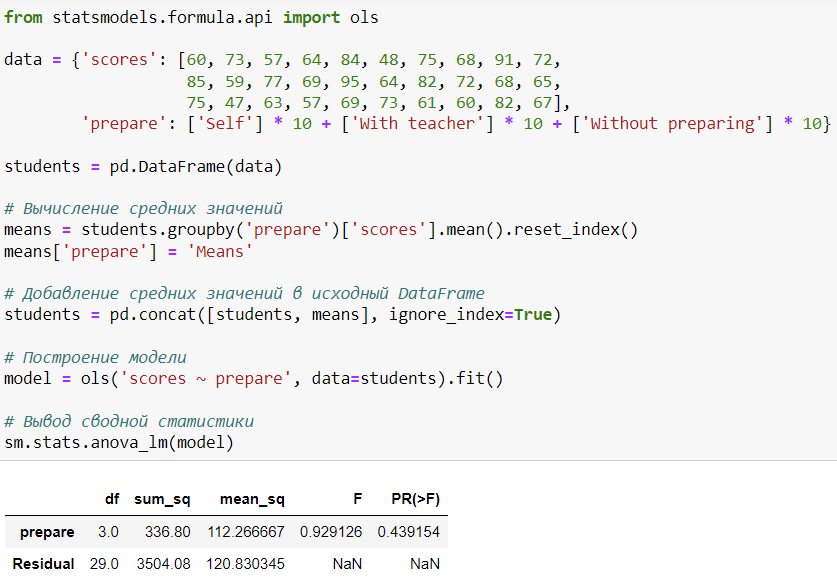






В нашей ситуации *Pr(>F) = 0.439 > 0.05 (уровень значимости).* Следовательно, мы не отвергаем нулевую гипотезу. Таким образом, с достаточно высокой степенью уверенности мы можем утверждать, что способ подготовки не оказал существенного влияния на результат контрольной.

**Python:**

****

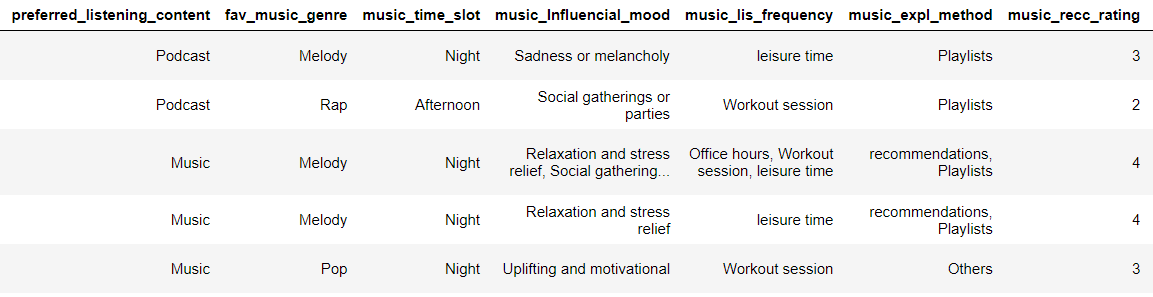
**Двухфакторный дисперсионный анализ** позволяет установить одновременное влияние двух факторов, а также взаимодействие между этими факторами.

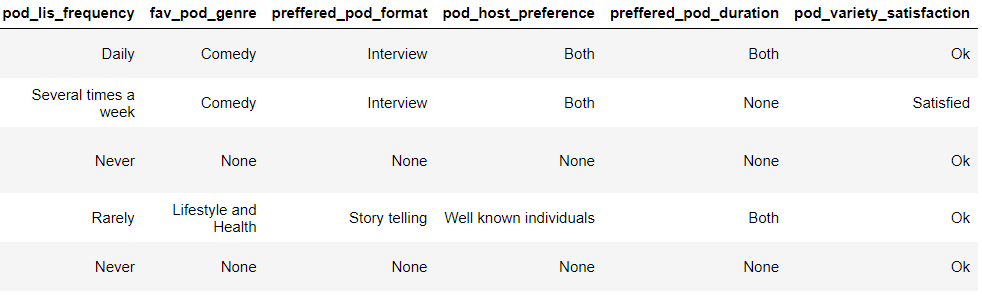
Исследуем датасет с информацией о пользователях Spotify. В нем содержится информация о возрасте, поле, любимом жанре музыки, плане подписки и других характеристиках пользователя. Ссылка на датасет: <https://www.kaggle.com/datasets/meeraajayakumar/spotify-user-behavior-dataset>





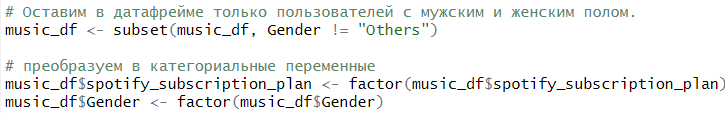




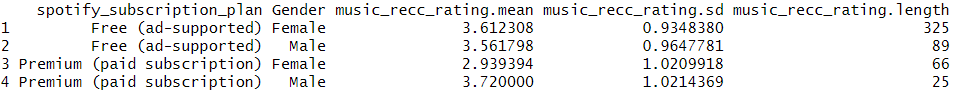


Выясним, влияет ли план подписки (**spotify\_subscription\_plan**) и пол (**Gender**) на то, какие оценки ставят пользователи на рекомендации от Spotify **(music\_recc\_rating)**.

**R:**

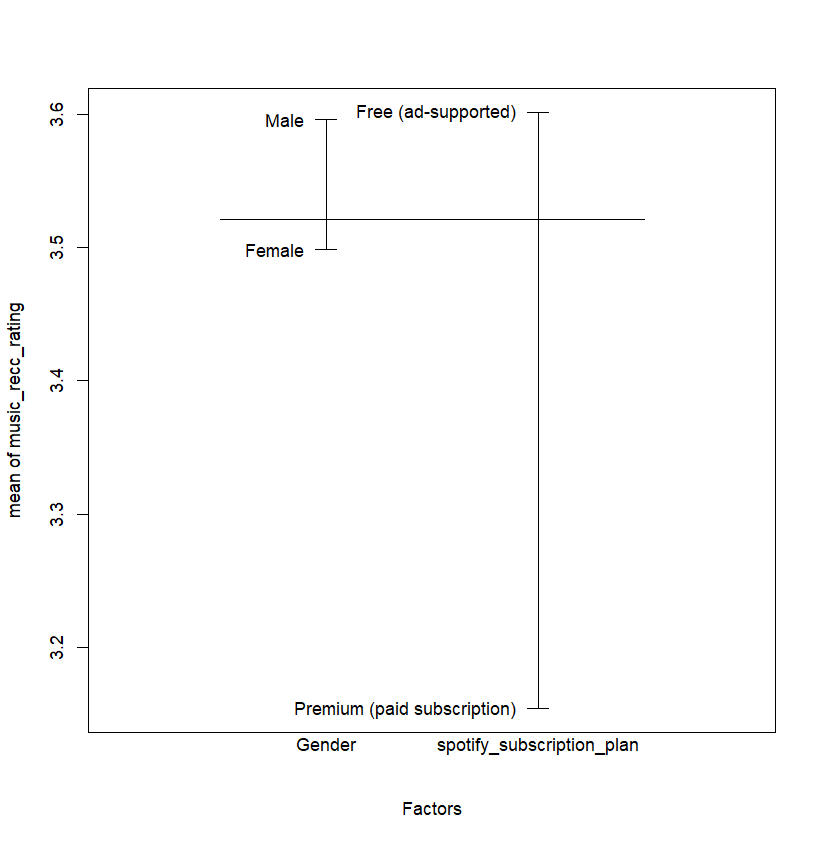
****

****

****

Видим, что у нас **несбалансированные** данные.

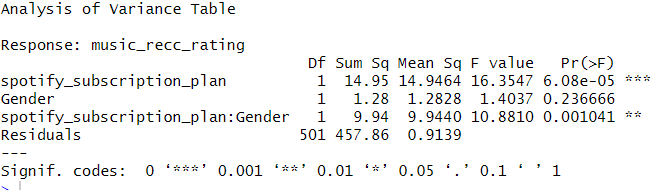




Из полученного графика видно, что наибольшая разница в средних оценках пользователей связана с планом подписки, тогда как эффект пола пользователя выражен в меньшей степени.

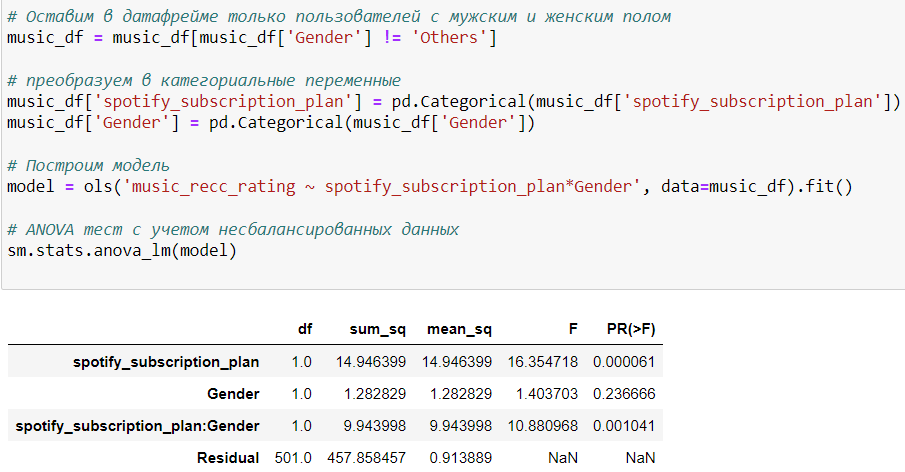
При анализе несбалансированных наборов данных, способ выполнения дисперсионного анализа, реализованный в функции aov(), будет давать смещенные оценки Р-значений. В таких случаях следует использовать функцию **lm()**.





Видим, что между оценками пользователей и планом подписки существует статистически значимая связь (*Pr(>F) = 6.08e-05*), тогда как влияние пола пользователя оказалось статистически незначимым (*Pr(>F) = 0.236*). Взаимодействие между планом подписки и полом пользователя статистически значимо (*Pr(>F) = 0.001041*).

**Python:**

****

**12. Подогнать регрессионные модели (в том числе, нелинейные) к**

**данным, а также оценить качество подобной аппроксимации**.

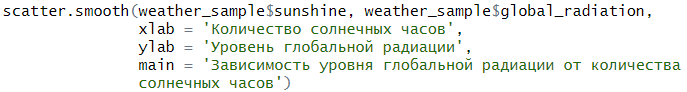
Вернемся к датасету с погодой в Лондоне. Подгоним простую модель **линейной регрессии** , используя количество солнечных часов (**sunshine**) в качестве переменной-предиктора и уровень глобальной радиации (global\_radiation) в качестве переменной ответа.

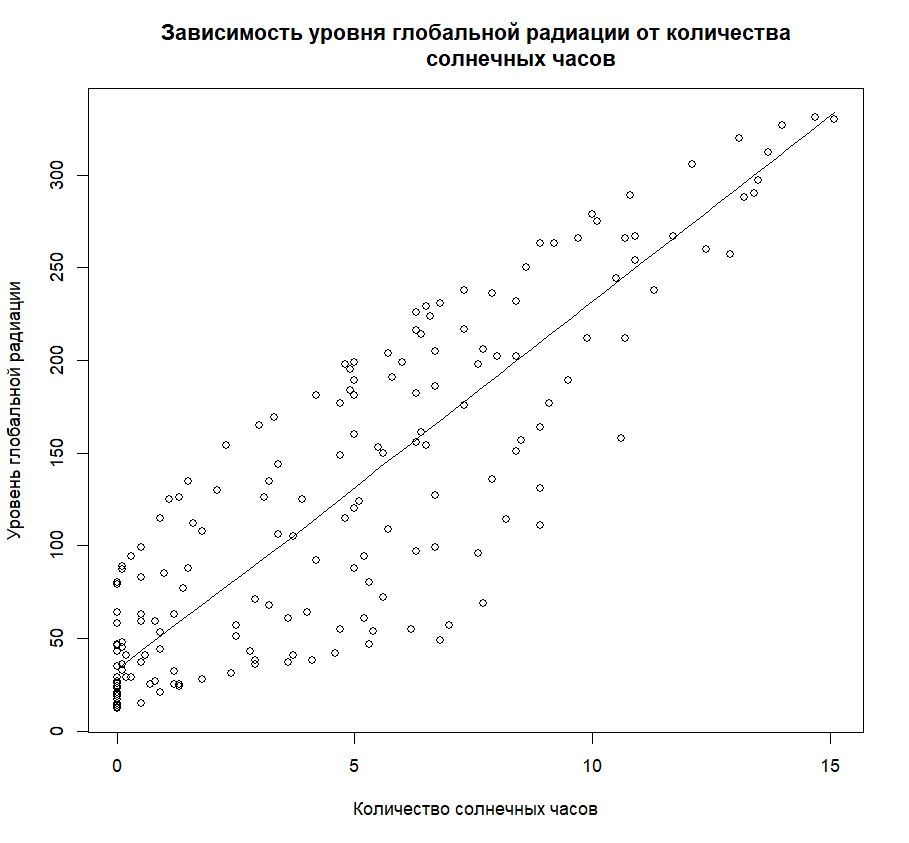
Выберем случайным образом 200 данных из датафрейма.

**R:**



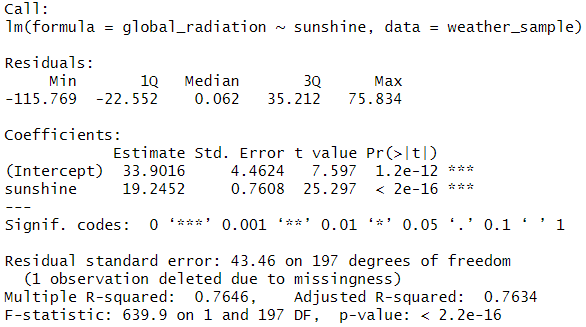
Убедимся, что связь между global\_radiation и sunshine примерно линейна.





Видим, что связь примерно линейна.





*Pr(>|t|) < 2e-16 << 0.05*. Значит, мы можем сказать, что существует статистически значимая связь между уровнем глобальной радиации и количеством солнечных часов.

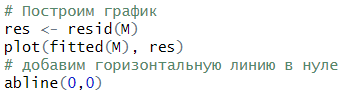
*Multiple R-squared*: это число говорит нам о том, какую процентную вариацию в уровне глобальной радииации можно объяснить количеством солнечных часов. В целом, чем больше значение R-квадрата регрессионной модели, тем лучше независимые переменные способны предсказать значение переменной отклика. В этом случае 76,46% вариации уровня глобальной радиации можно объяснить количеством солнечных часов.

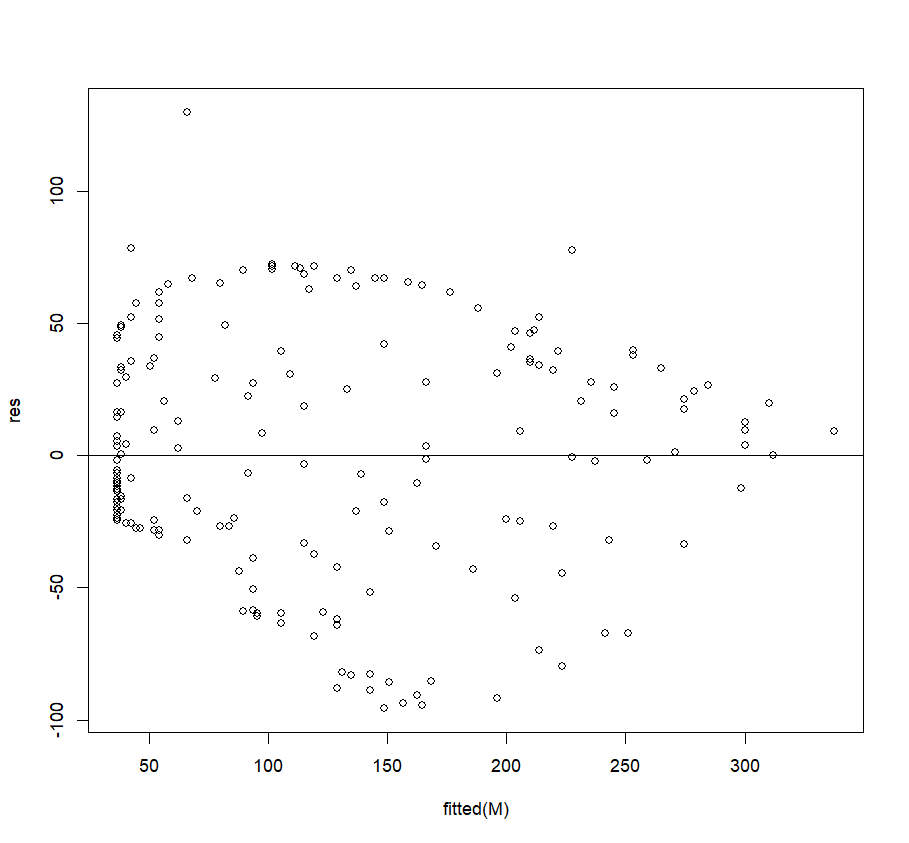
После того, как мы подогнали простую модель линейной регрессии к данным, последним шагом является создание остаточных графиков.

Одно из ключевых предположений линейной регрессии состоит в том, что остатки регрессионной модели примерно нормально распределены и гомоскедастичны на каждом уровне объясняющей переменной (предиктора). Если эти допущения нарушаются, то результаты нашей регрессионной модели могут вводить в заблуждение или быть ненадежными.

Чтобы убедиться, что эти предположения выполняются, мы можем построить следующие остаточные графики:

**График остаточных и подобранных значений**: этот график полезен для подтверждения гомоскедастичности. На оси X отображаются подогнанные значения, а на оси Y — остатки. Пока остатки кажутся распределенными по диаграмме вокруг нулевого значения, мы можем предположить, что гомоскедастичность не нарушена:

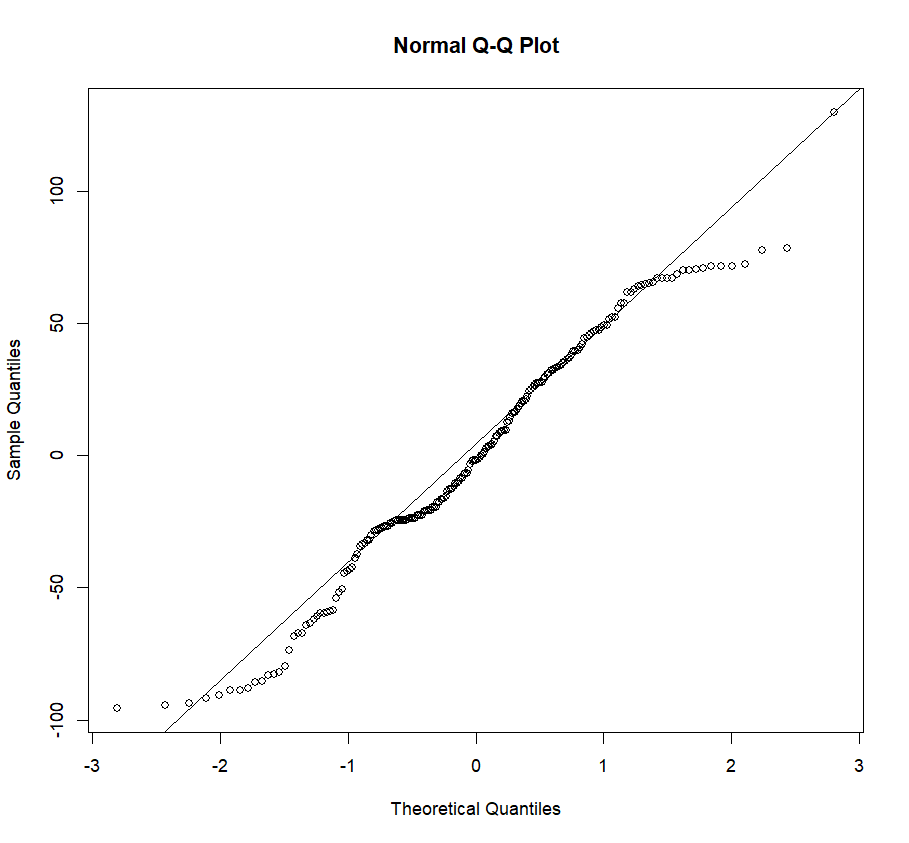




Остатки кажутся случайным образом разбросанными вокруг нуля и не демонстрируют каких-либо заметных закономерностей, так что гомоскедастичность не нарушена.

**График QQ:** этот график полезен для определения того, следуют ли остатки нормальному распределению. Если значения данных на графике падают примерно по прямой линии под углом 45 градусов, то данные распределяются нормально:

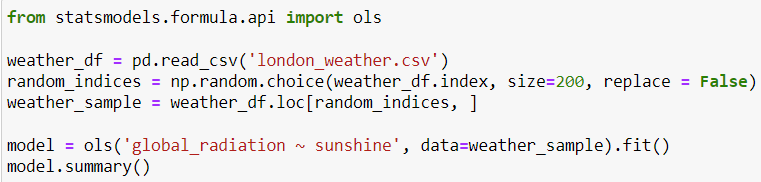


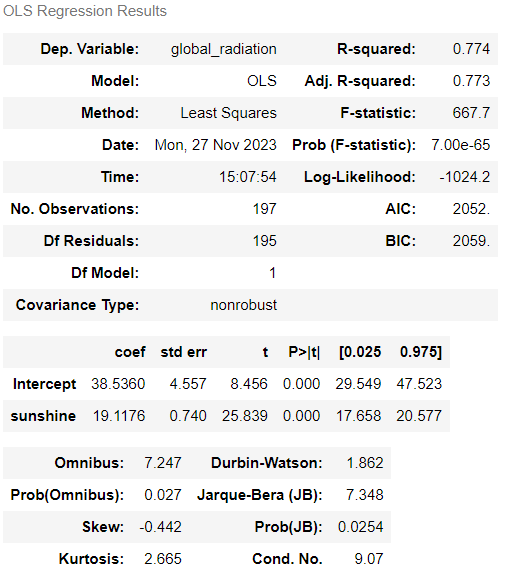


Остатки немного отклоняются от 45-градусной линии, но не настолько, чтобы вызывать серьезное беспокойство. Можно считать, что условие нормальности выполнено.

Поскольку остатки нормально распределены и гомоскедастичны, мы проверили выполнение допущений модели простой линейной регрессии. Таким образом, **вывод нашей модели надежен**.

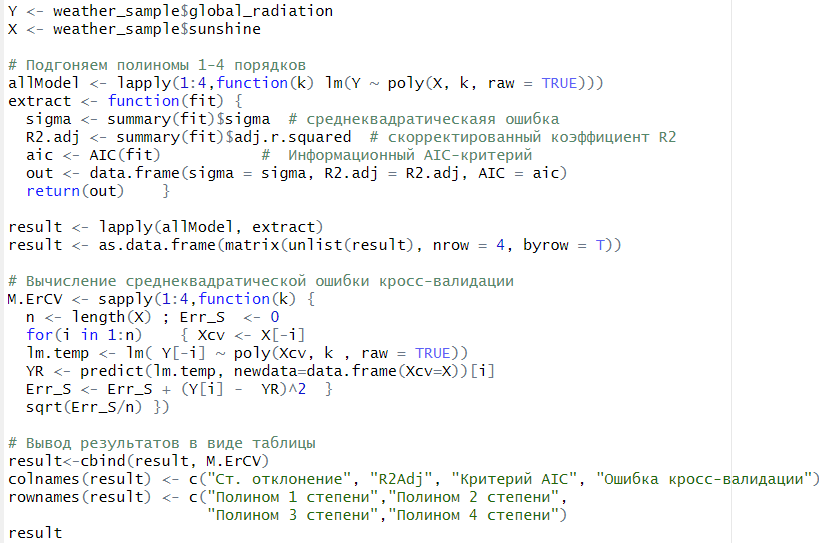
**Python:**

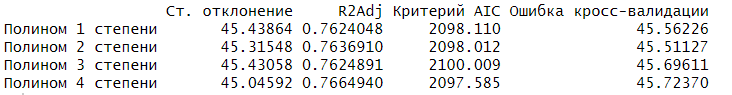




Теперь воспользуемся **полиномиальной** регрессией.

**R:**





Видим, что результаты полиномов разных степеней мало отличаются.

Кроме того, видим, что все показатели полиномиальной и построенной раннее линейной регрессии практически идентичны.