**Московский государственный технический университет им. Н.Э.Баумана**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Курсовая работа по дисциплине**

**«Технологии Машинного Обучения»**

Пояснительная записка

(вид документа)

писчая бумага

(вид носителя)

13

(количество листов)

|  |  |
| --- | --- |
| Исполнитель: |  |
| студентка группы РТ5-71 |  |
| Каркавина А. А. | "\_\_\_"\_\_\_\_\_\_\_\_\_2019 г. |

|  |  |
| --- | --- |
| Утверждаю: |  |
| Гапанюк Ю.Е. | "\_\_"\_\_\_\_\_\_\_\_\_\_\_\_\_2019 г. |

Москва - 2019

**Оглавление**

1. Задание 2
2. Введение 2
3. Основная часть 3
4. Заключение 14
5. Источники 14

**Задание:**

1. Поиск и выбор набора данных для построения моделей машинного обучения. На основе выбранного набора данных студент должен построить модели машинного обучения для решения или задачи классификации, или задачи регрессии.

2. Проведение разведочного анализа данных. Построение графиков, необходимых для понимания структуры данных. Анализ и заполнение пропусков в данных.

3. Выбор признаков, подходящих для построения моделей. Кодирование категориальных признаков Масштабирование данных. Формирование вспомогательных признаков, улучшающих качество моделей.

4. Проведение корреляционного анализа данных. Формирование промежуточных выводов о возможности построения моделей машинного обучения. В зависимости от набора данных, порядок выполнения пунктов 2, 3, 4 может быть изменен.

5. Выбор метрик для последующей оценки качества моделей. Необходимо выбрать не менее трех метрик и обосновать выбор.

6. Выбор наиболее подходящих моделей для решения задачи классификации или регрессии. Необходимо использовать не менее пяти моделей, две из которых должны быть ансамблевыми.

7. Формирование обучающей и тестовой выборок на основе исходного набора данных.

8. Построение базового решения (baseline) для выбранных моделей без подбора гиперпараметров. Производится обучение моделей на основе обучающей выборки и оценка качества моделей на основе тестовой выборки.

9. Подбор гиперпараметров для выбранных моделей. Рекомендуется использовать методы кросс-валидации. В зависимости от используемой библиотеки можно применять функцию GridSearchCV, использовать перебор параметров в цикле, или использовать другие методы.

10. Повторение пункта 8 для найденных оптимальных значений гиперпараметров. Сравнение качества полученных моделей с качеством baseline-моделей.

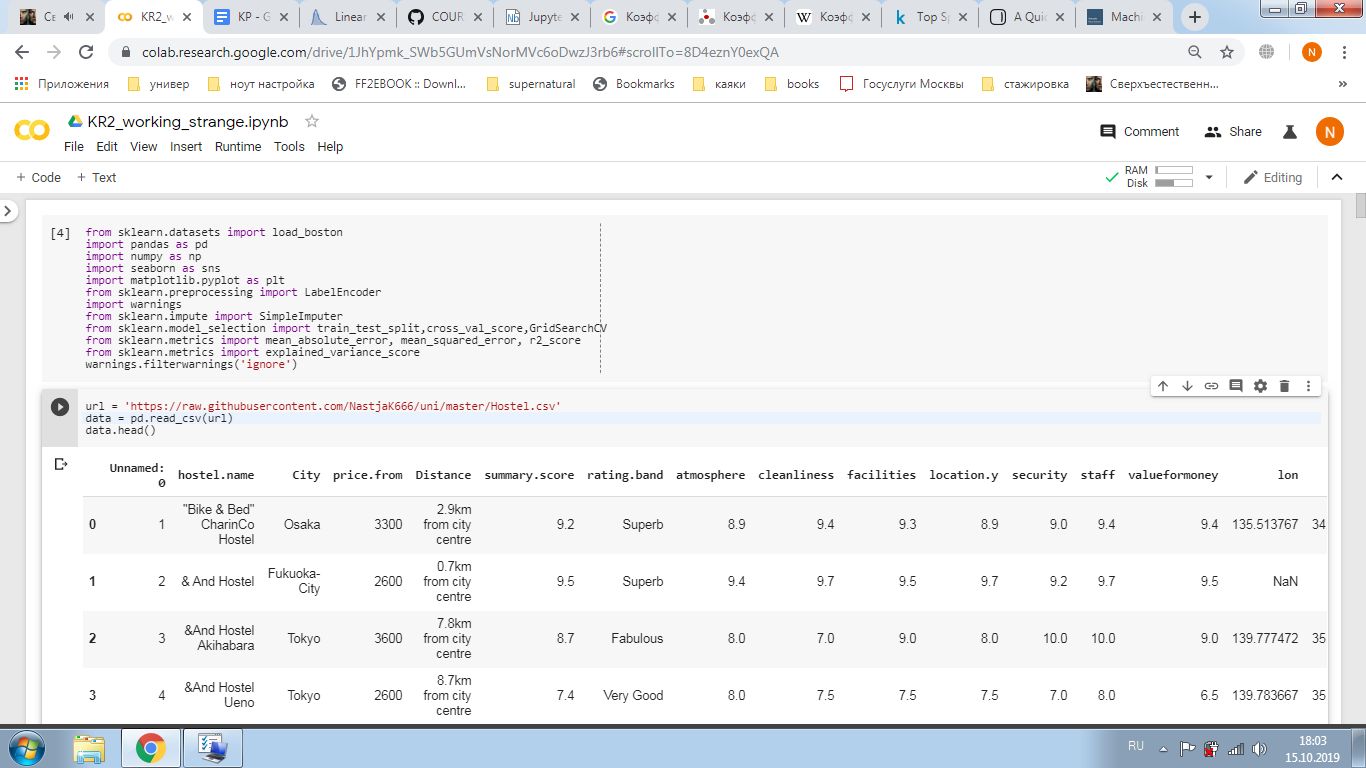
11. Формирование выводов о качестве построенных моделей на основе выбранных метрик. Результаты сравнения качества рекомендуется отобразить в виде графиков и сделать выводы в форме текстового описания. Рекомендуется построение графиков обучения и валидации, влияния значений гиперпарметров на качество моделей и т.д.

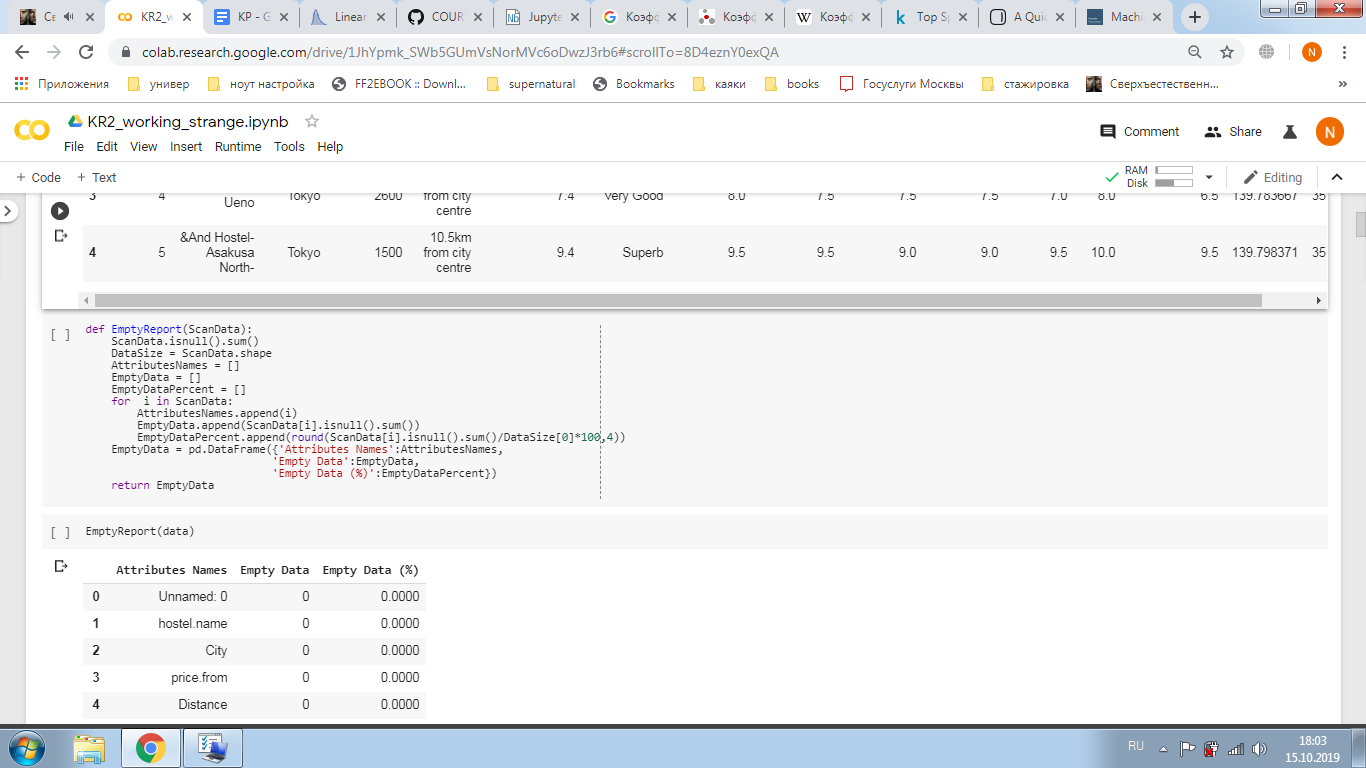
**Введение:**

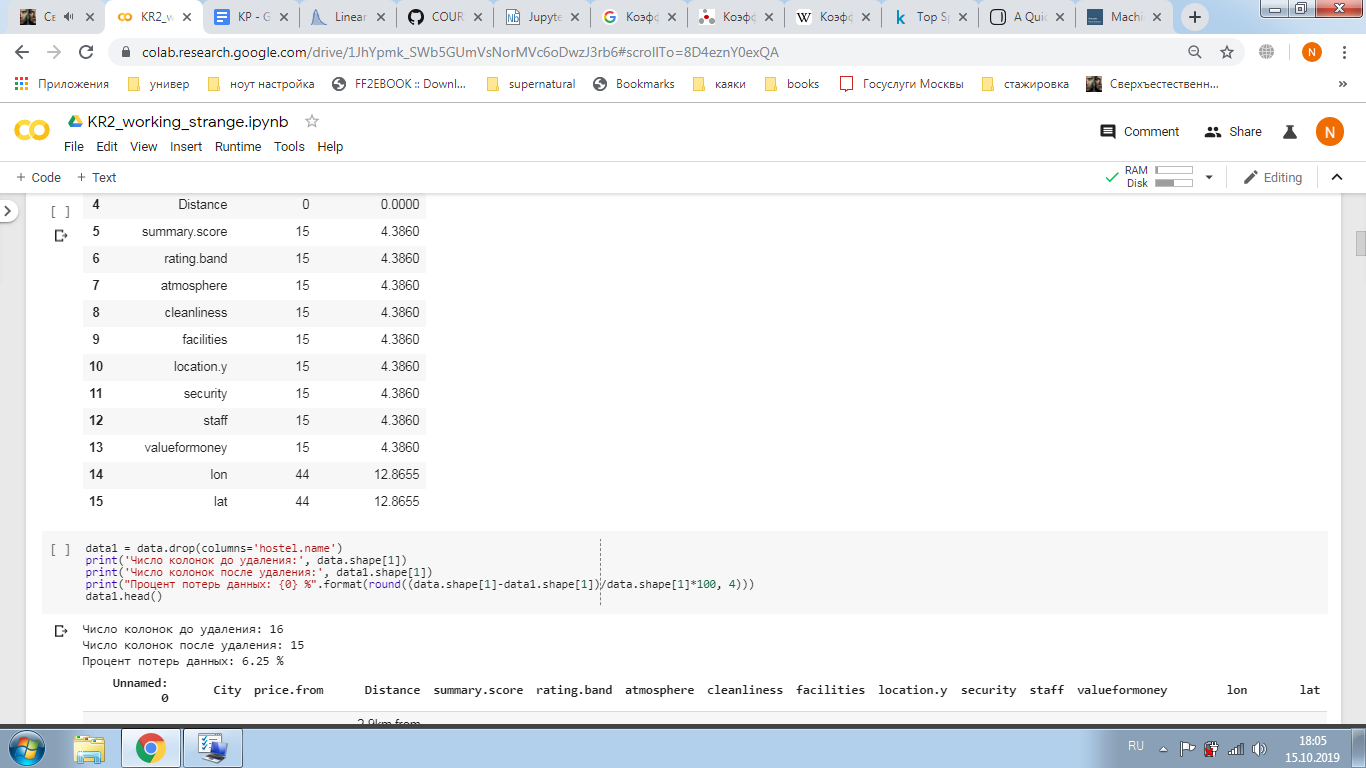
Курсовая работа – самостоятельная часть учебной дисциплины «Технологии машинного обучения» – учебная и практическая исследовательская студенческая работа, направленная на решение комплексной задачи машинного обучения. Результатом курсовой работы является отчет, содержащий описания моделей, тексты программ и результаты экспериментов.

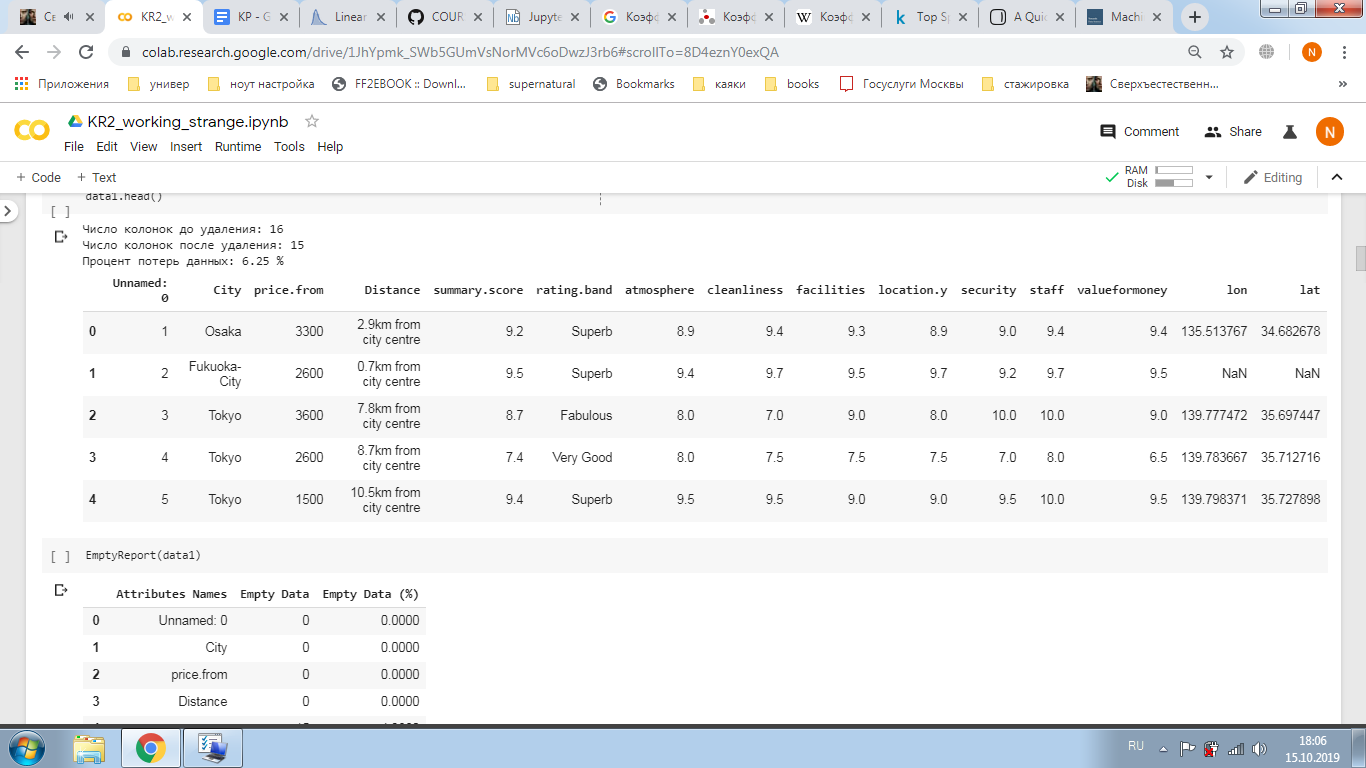
Курсовая работа опирается на знания, умения и владения, полученные студентом в рамках лекций и лабораторных работ по дисциплине.

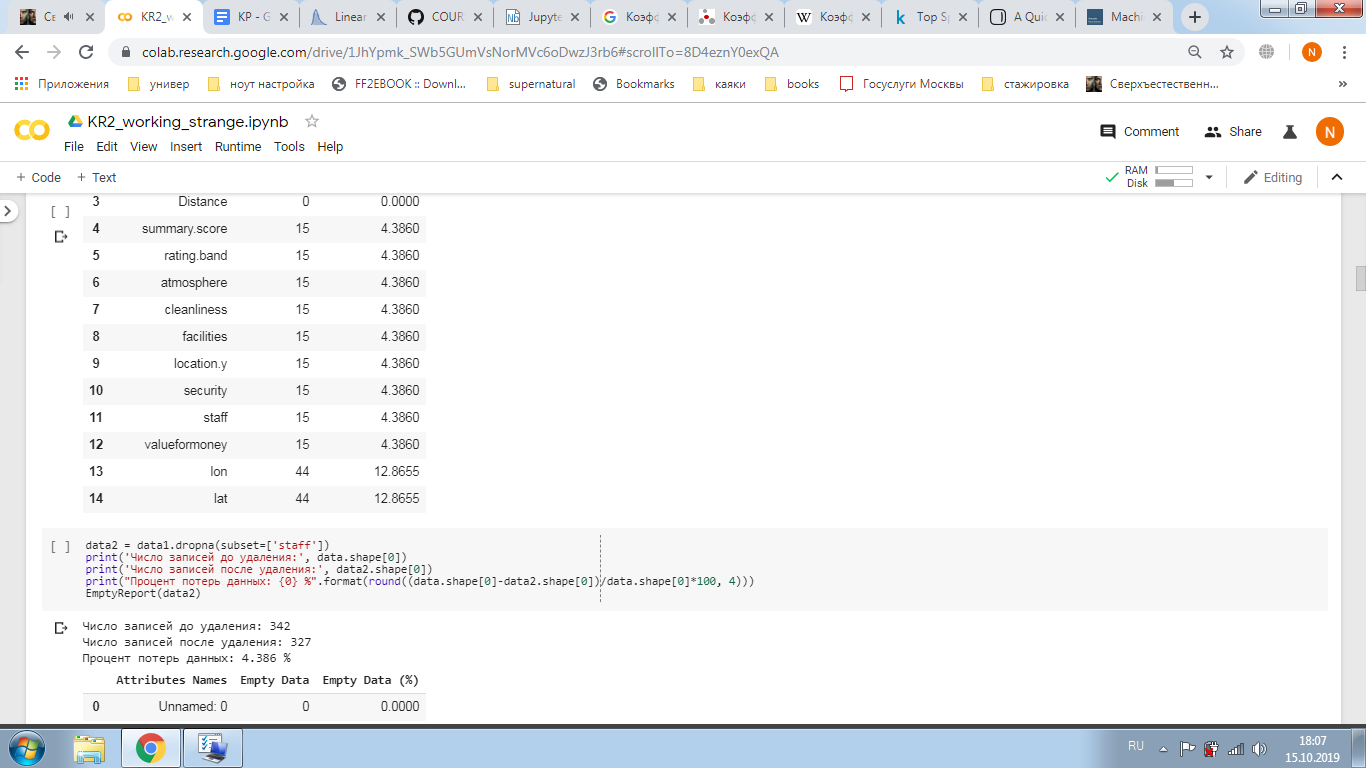
Основная часть:

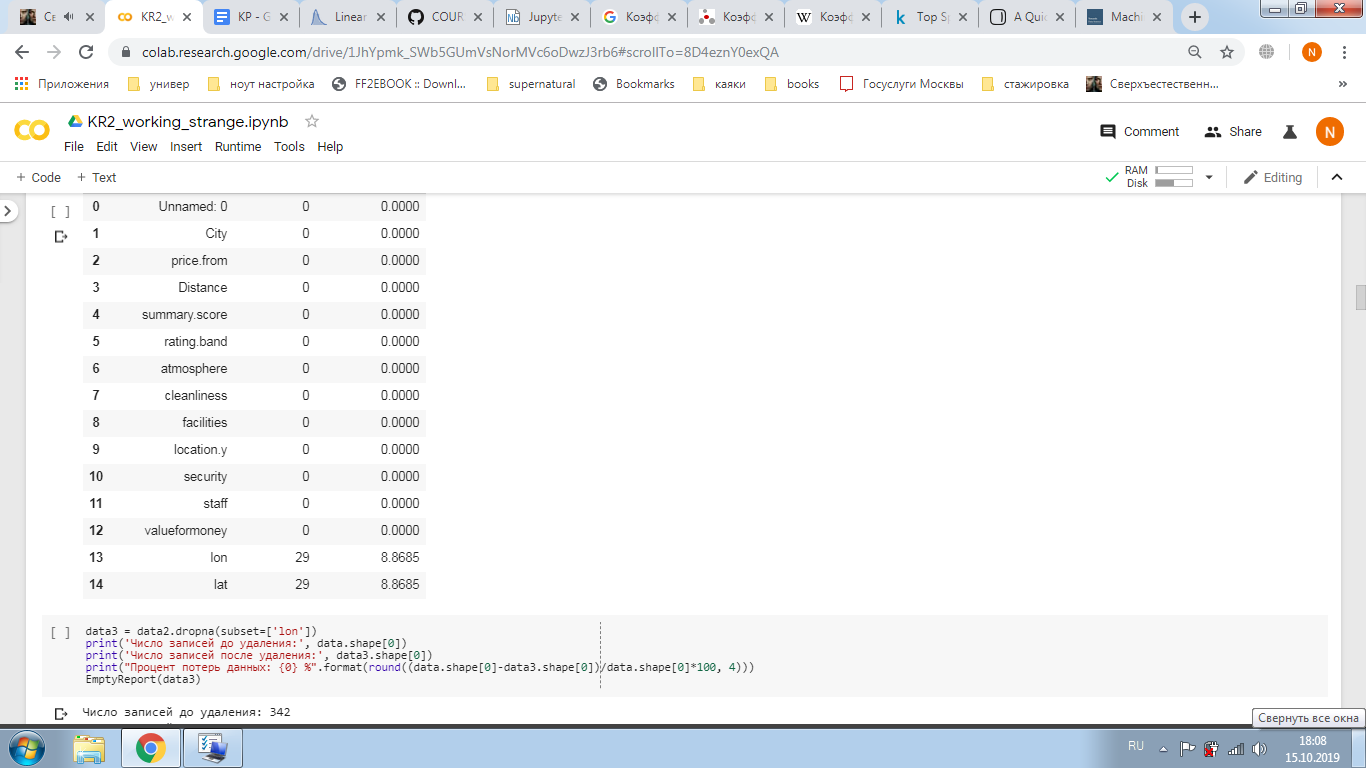


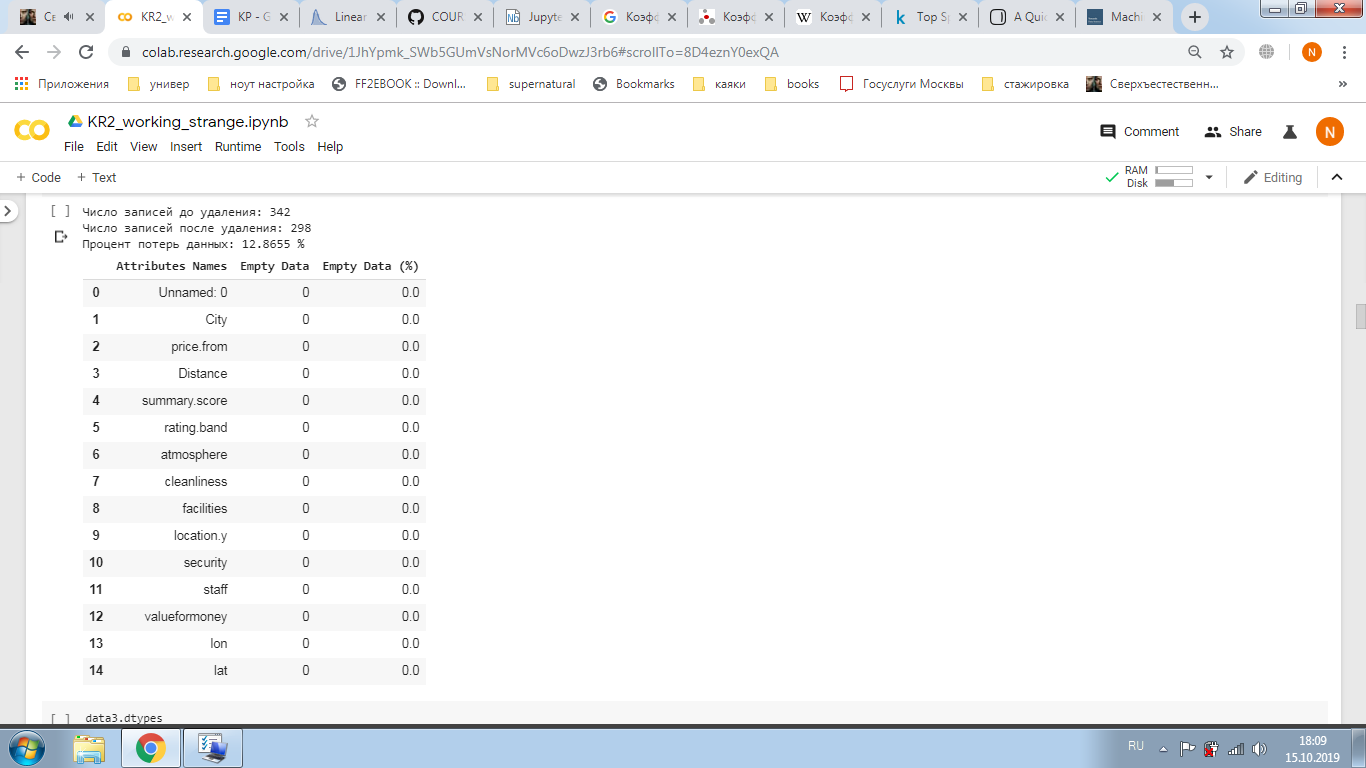


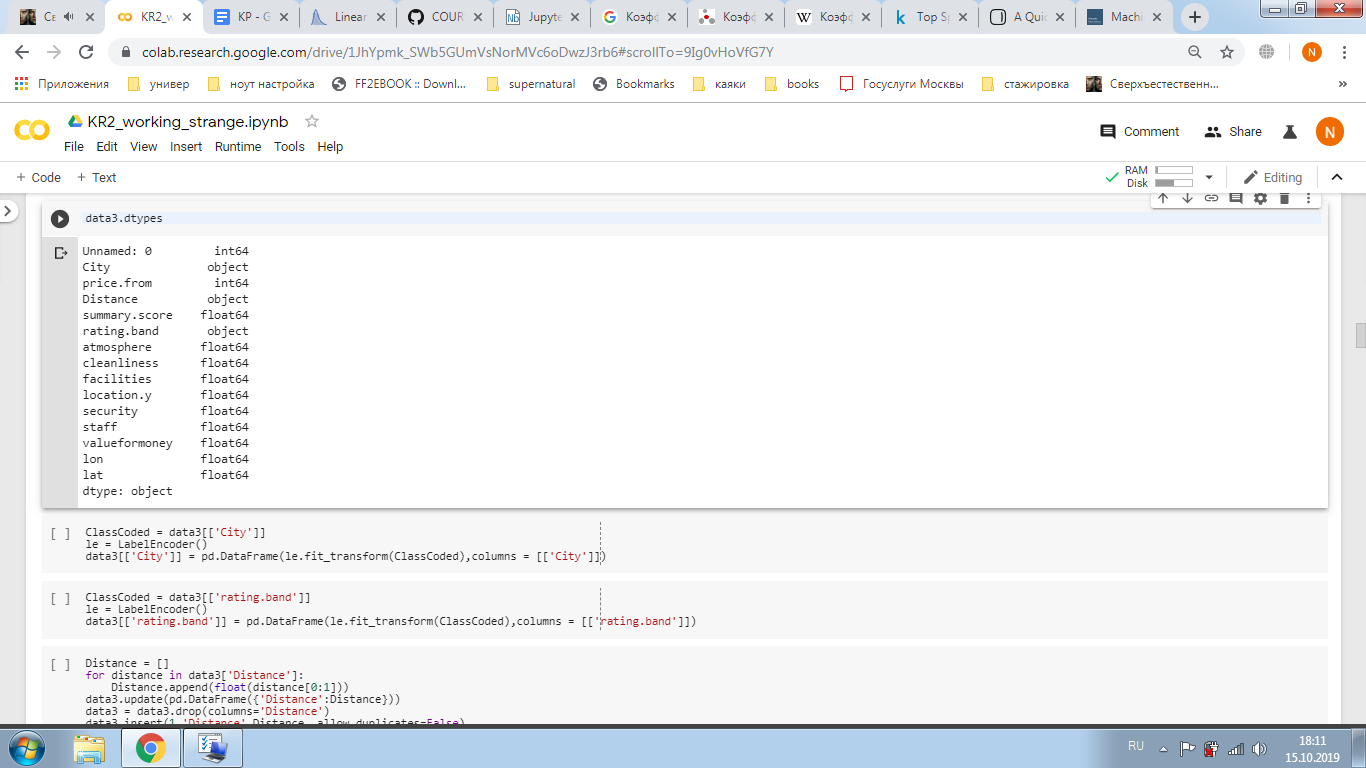


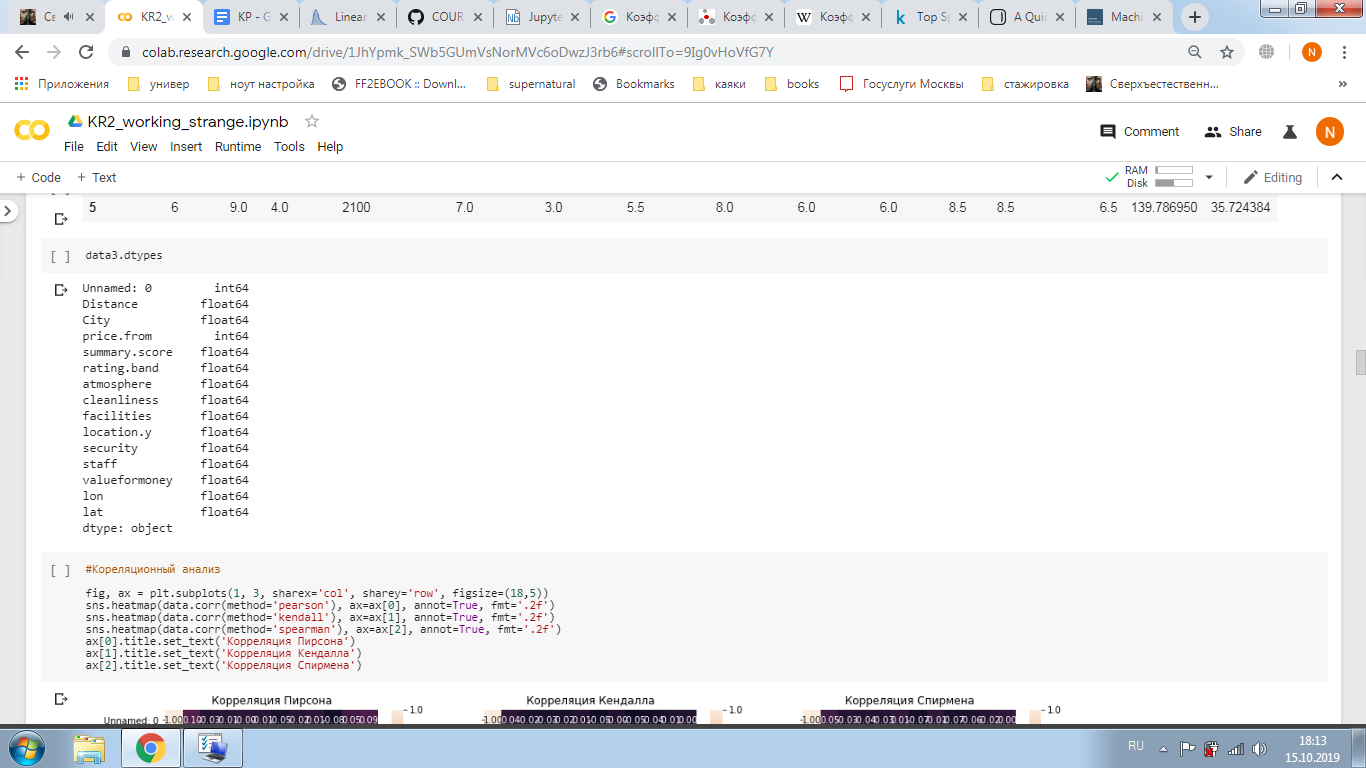
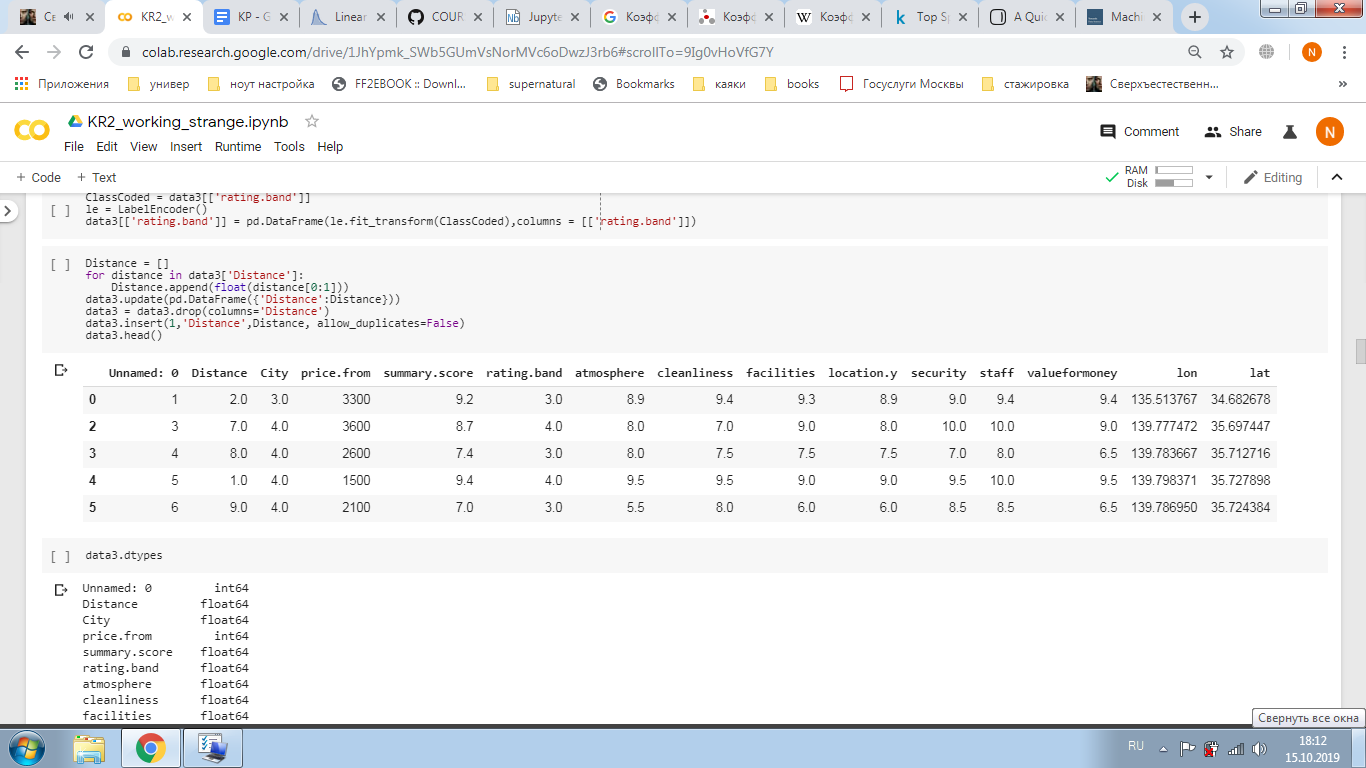


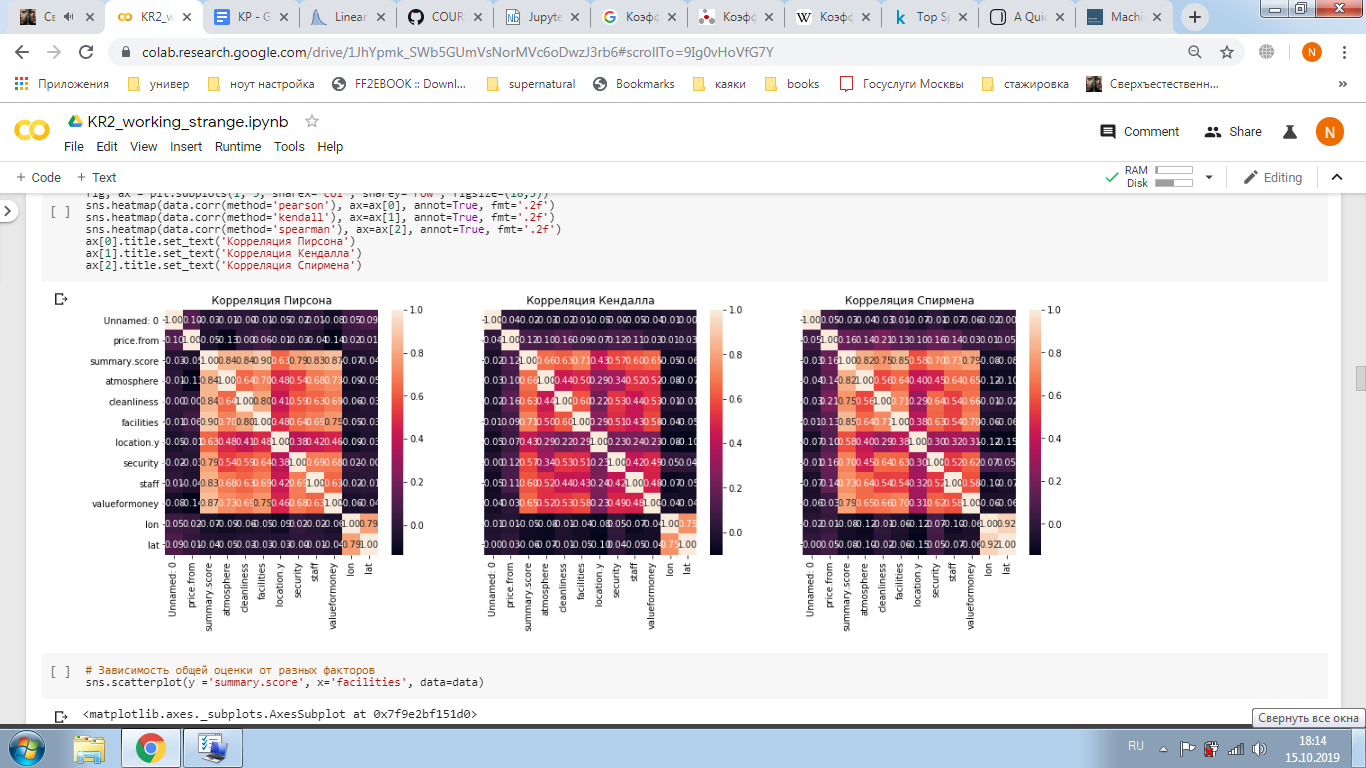


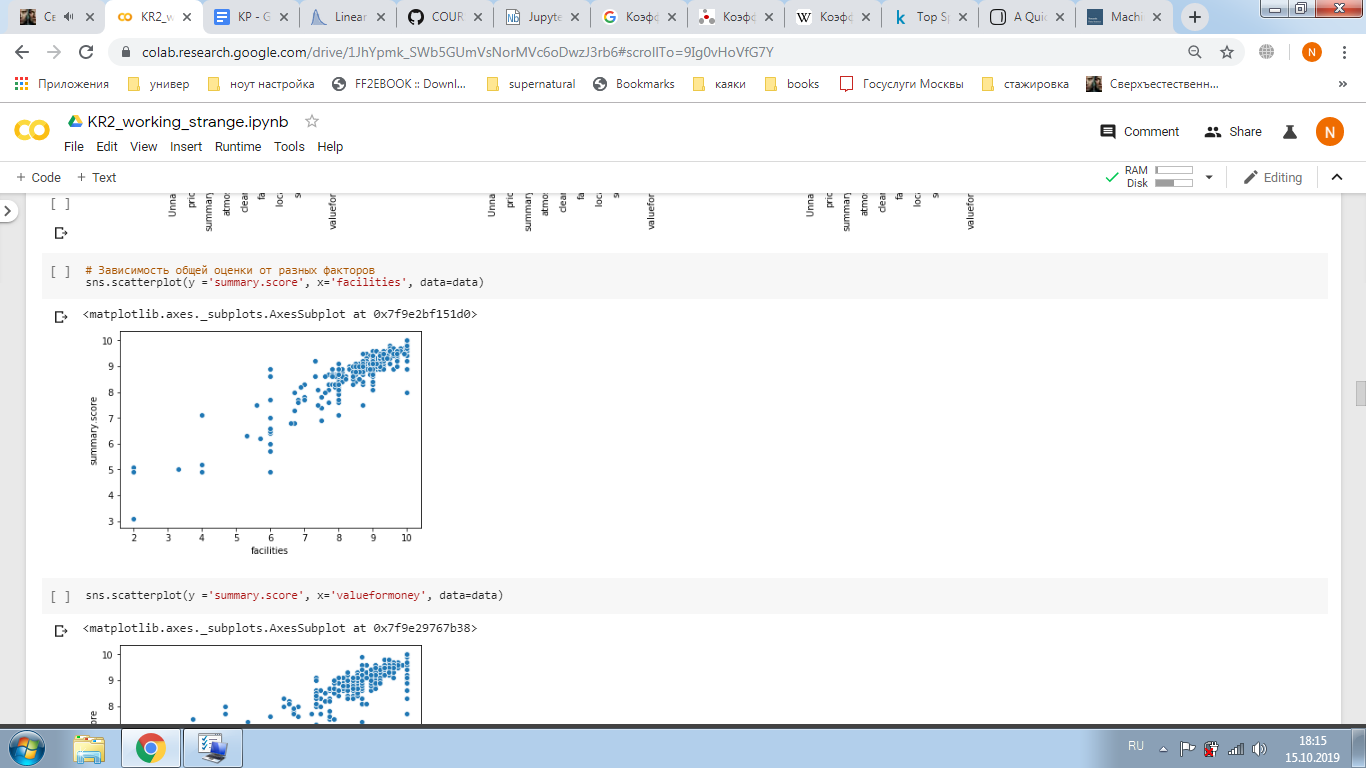


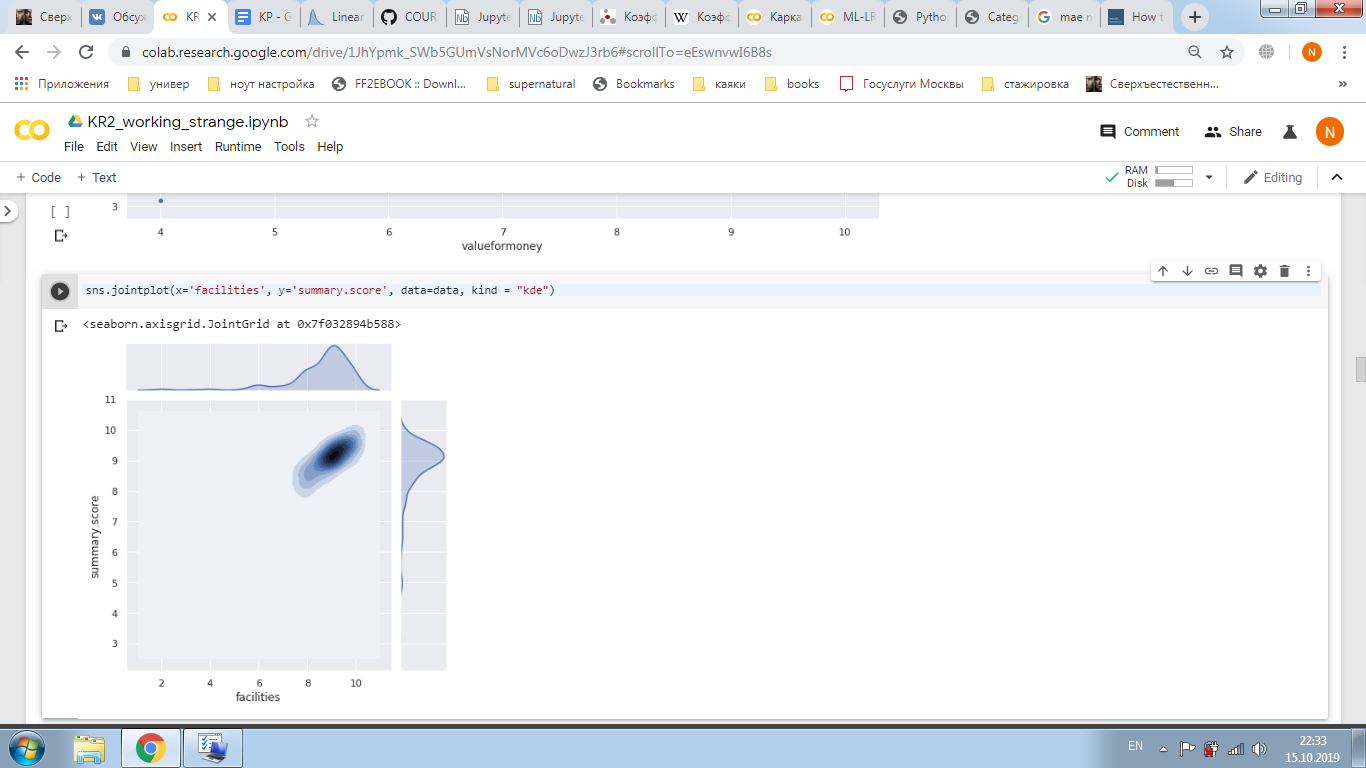


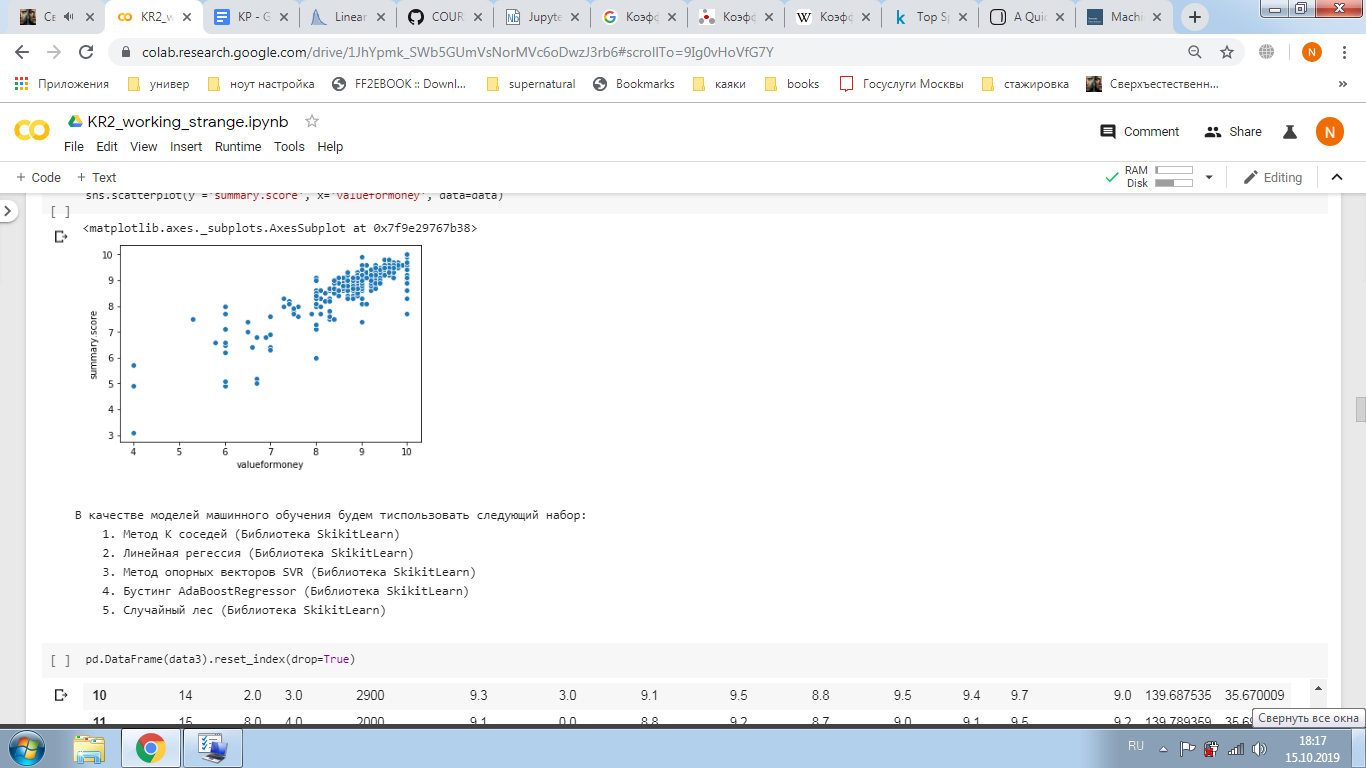


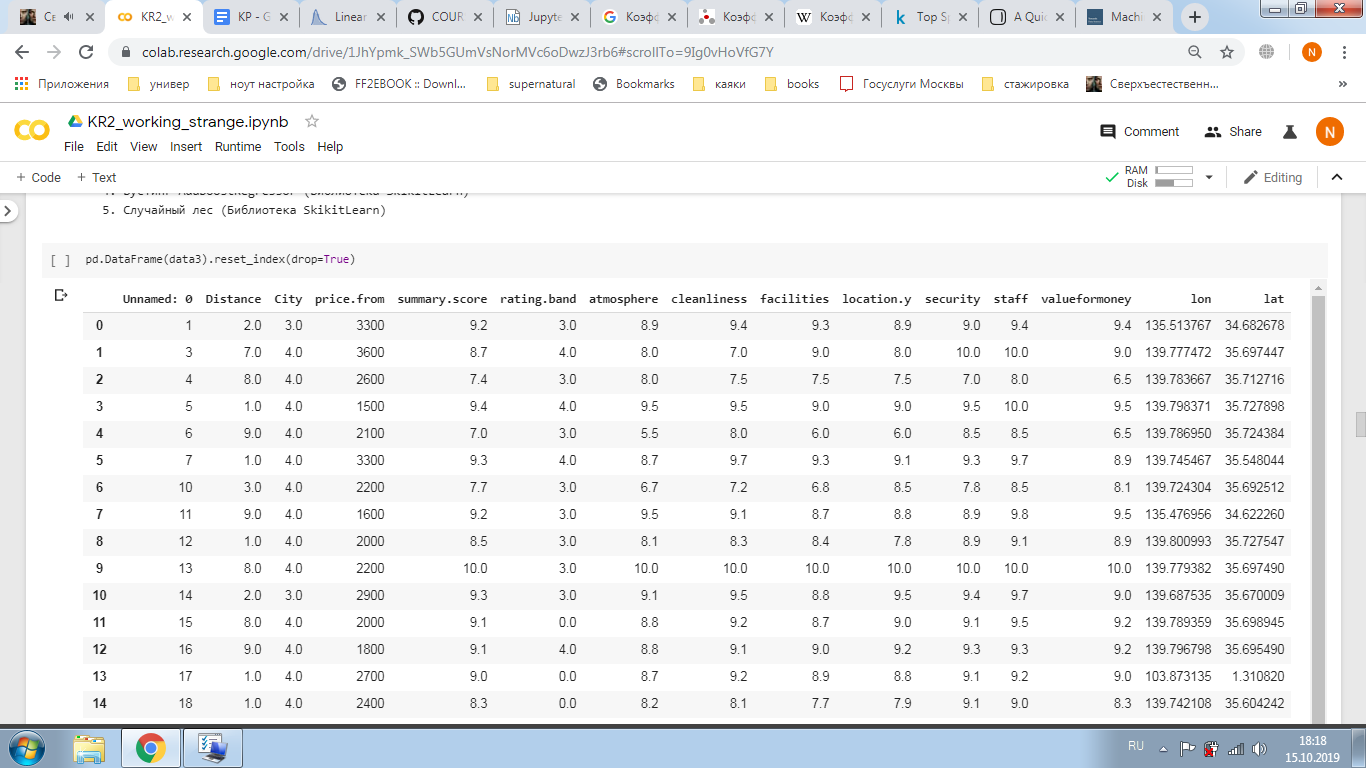


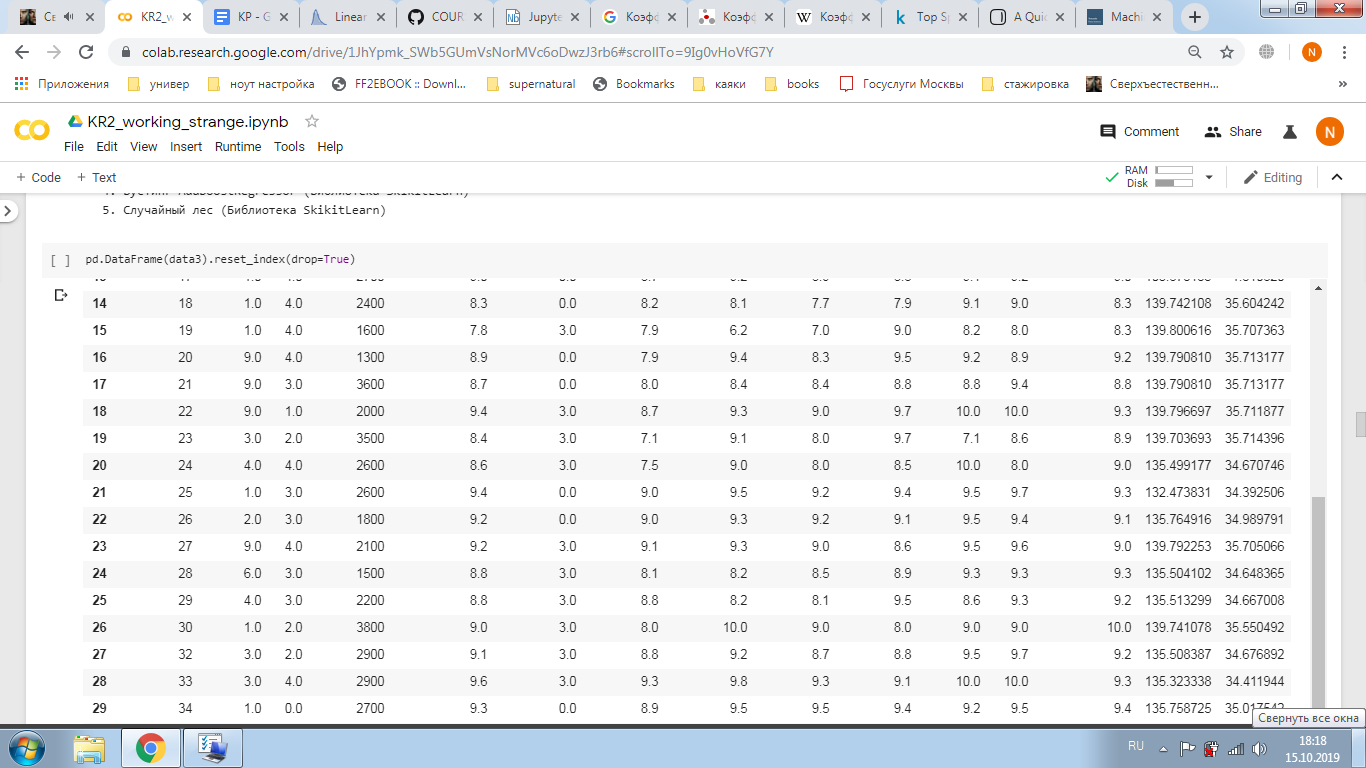


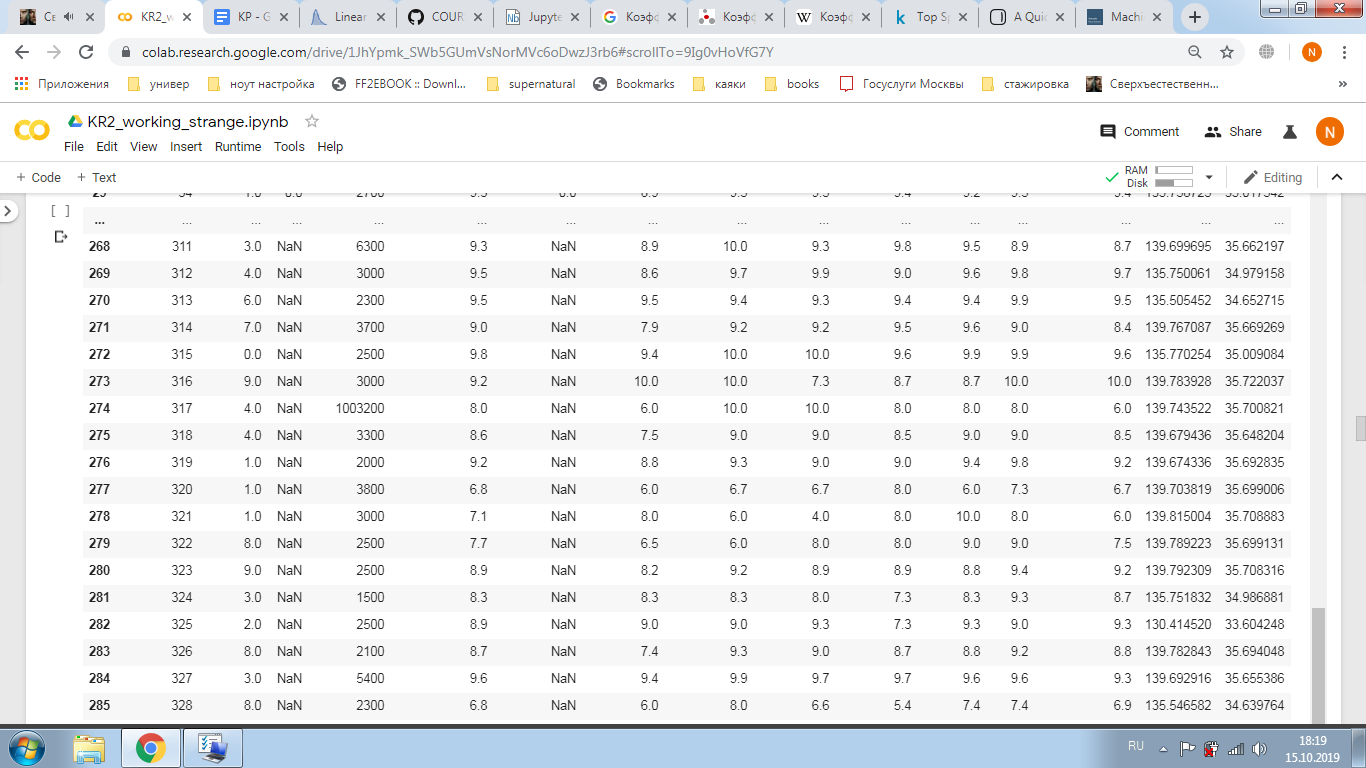


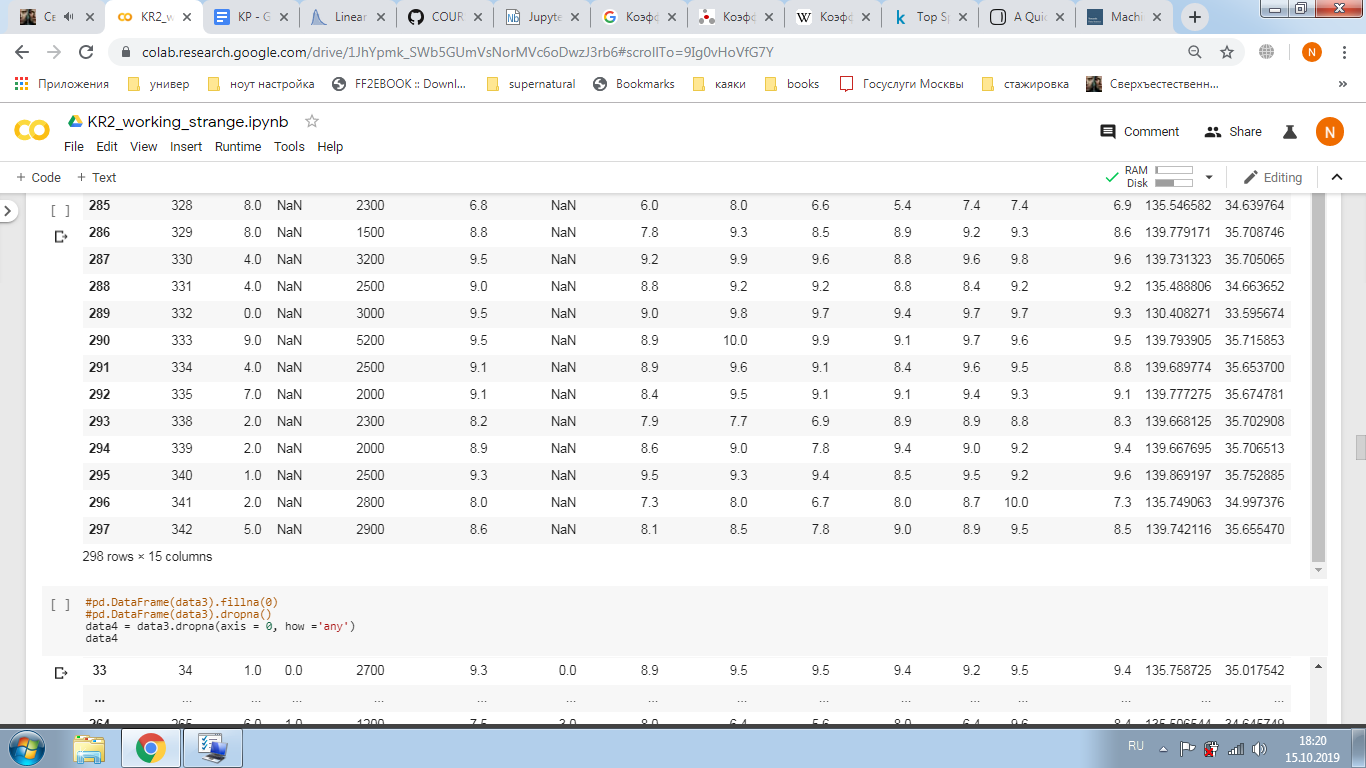


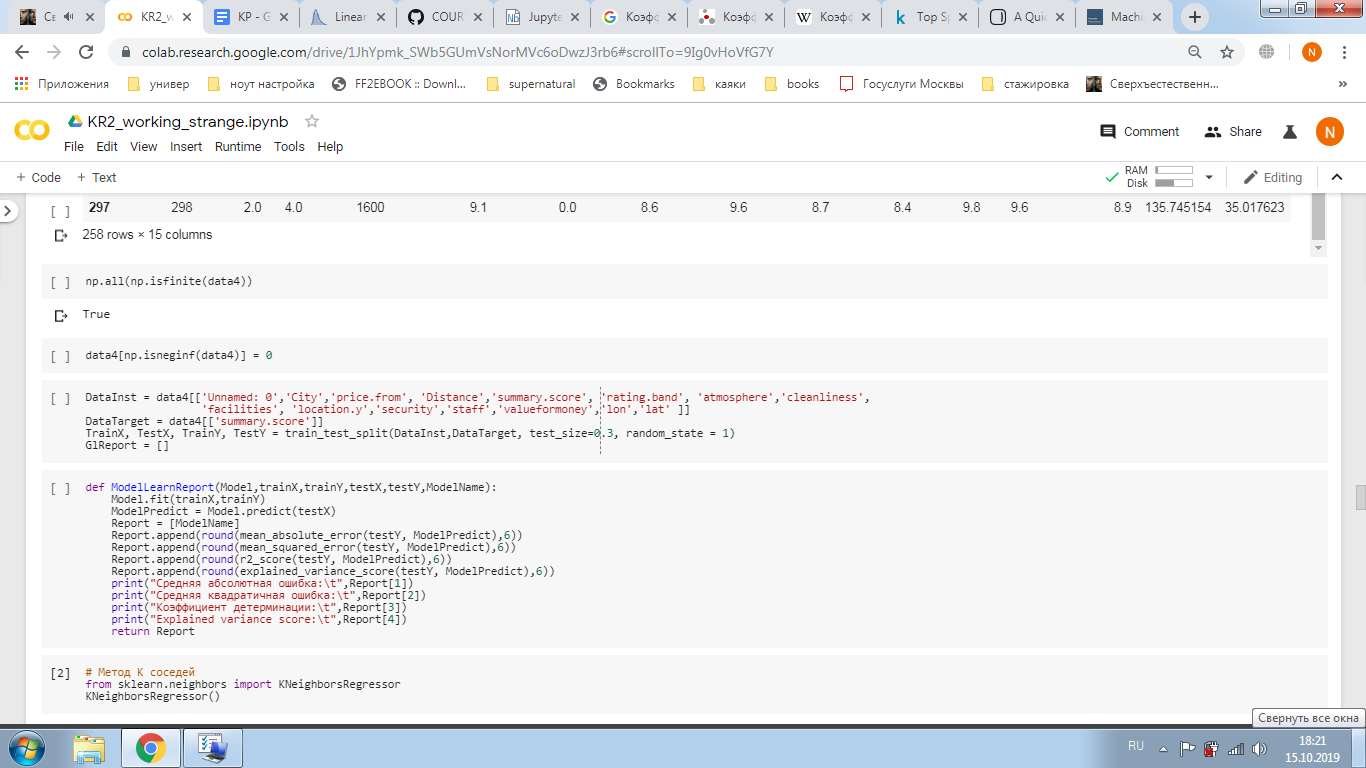


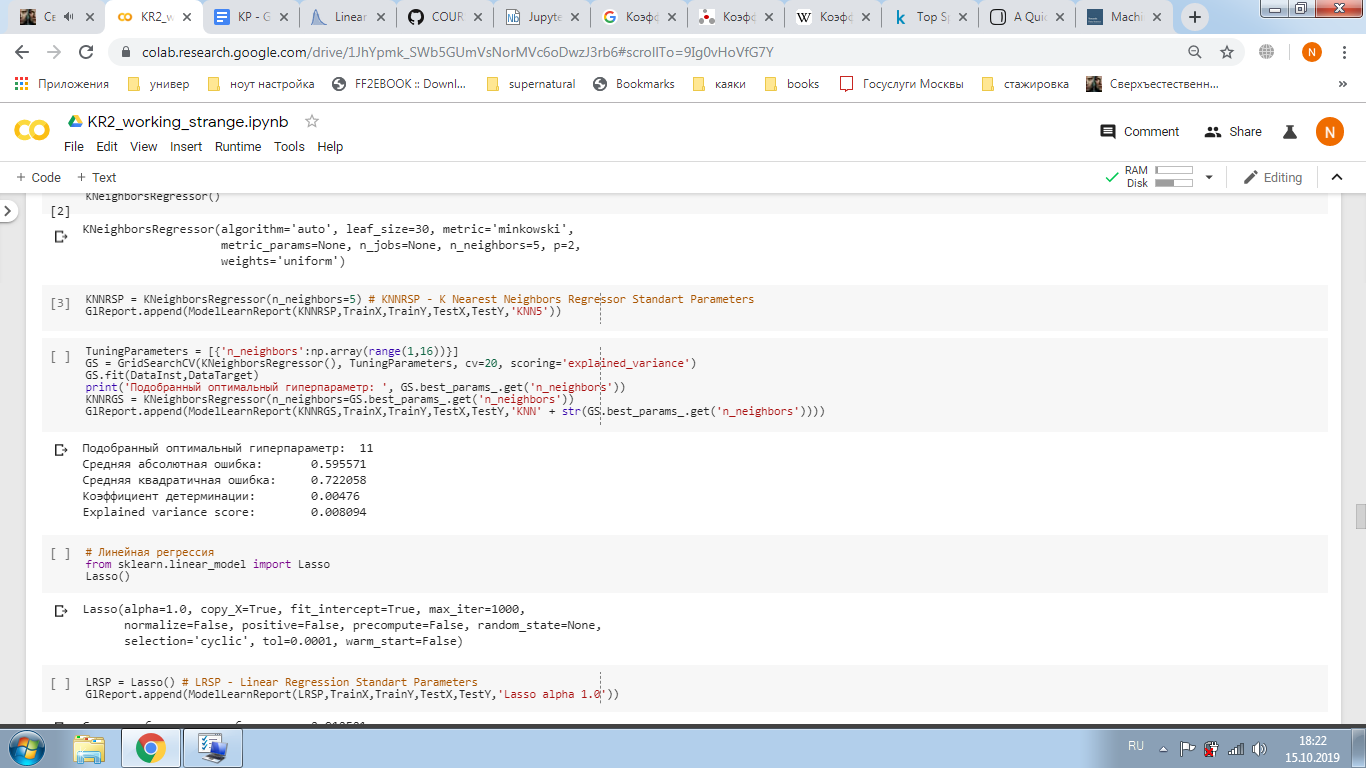




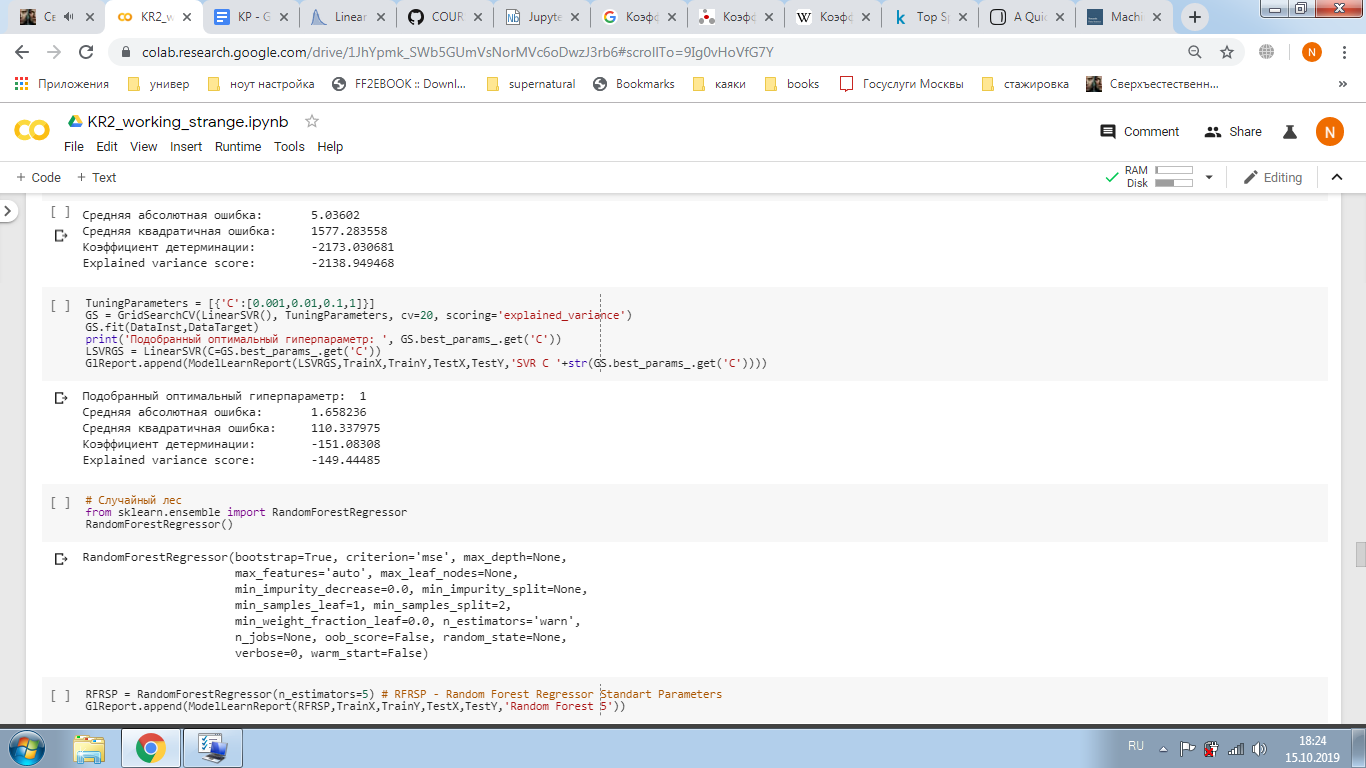


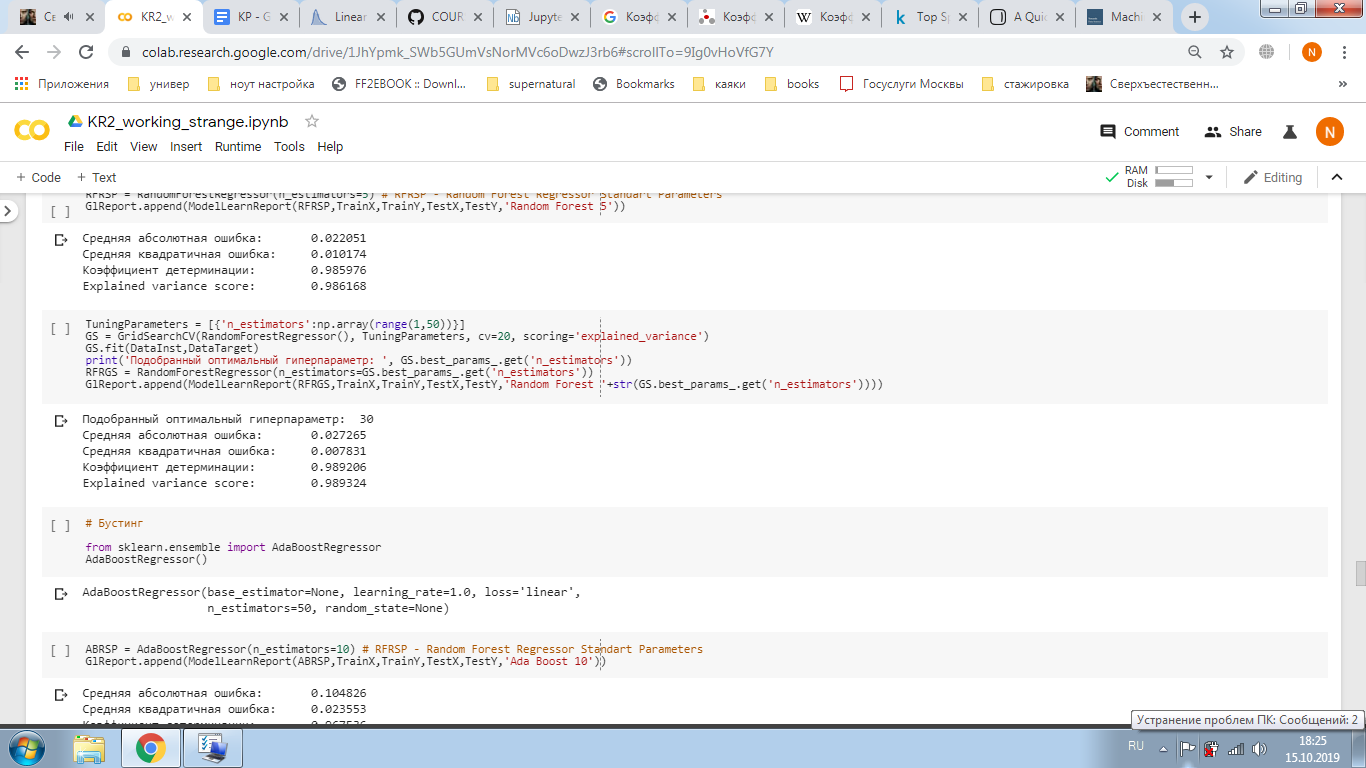


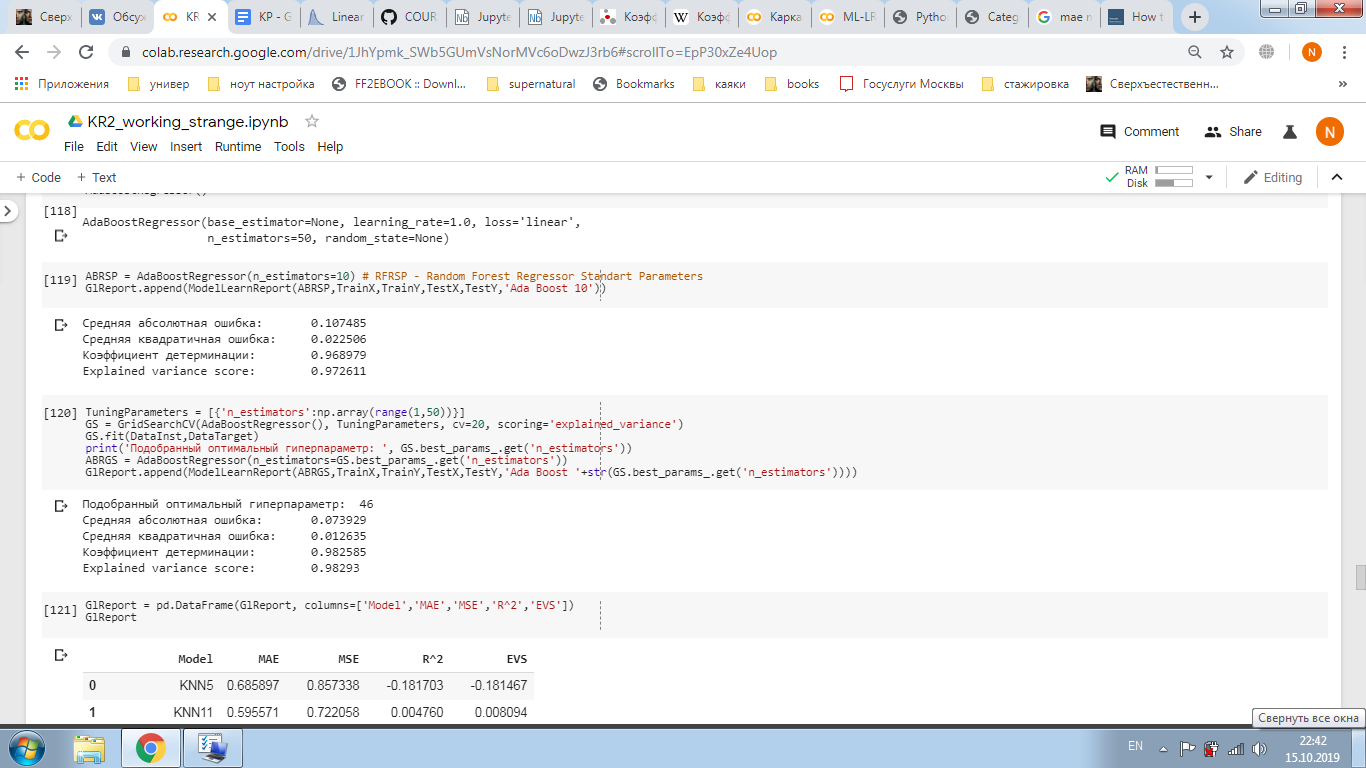


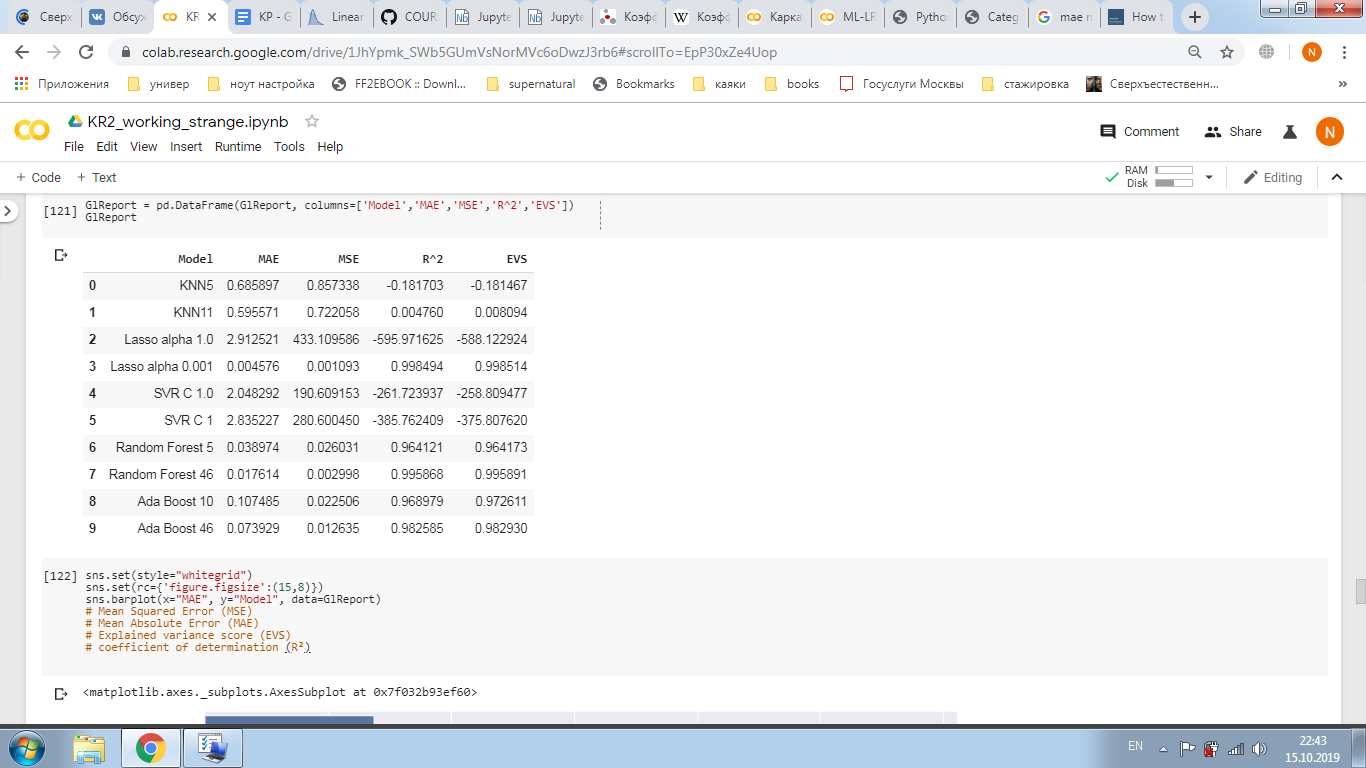


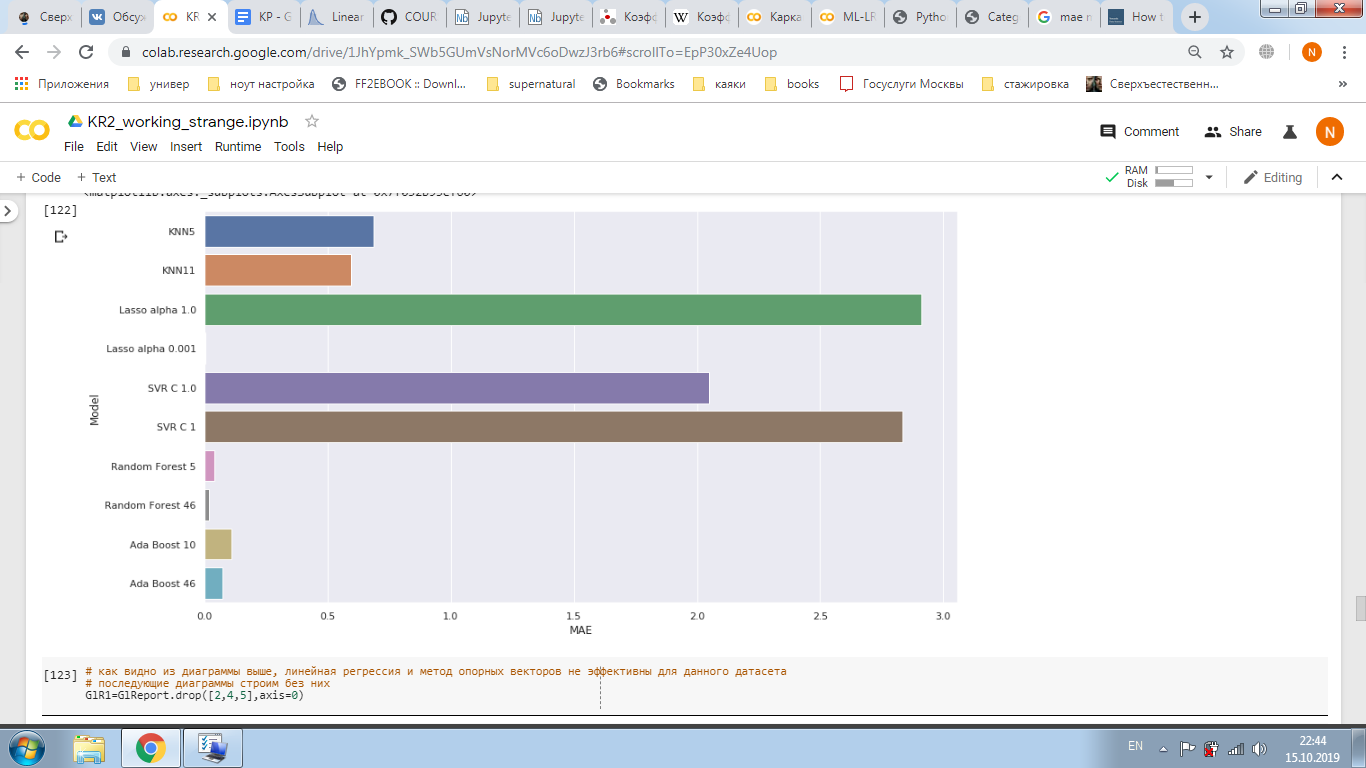


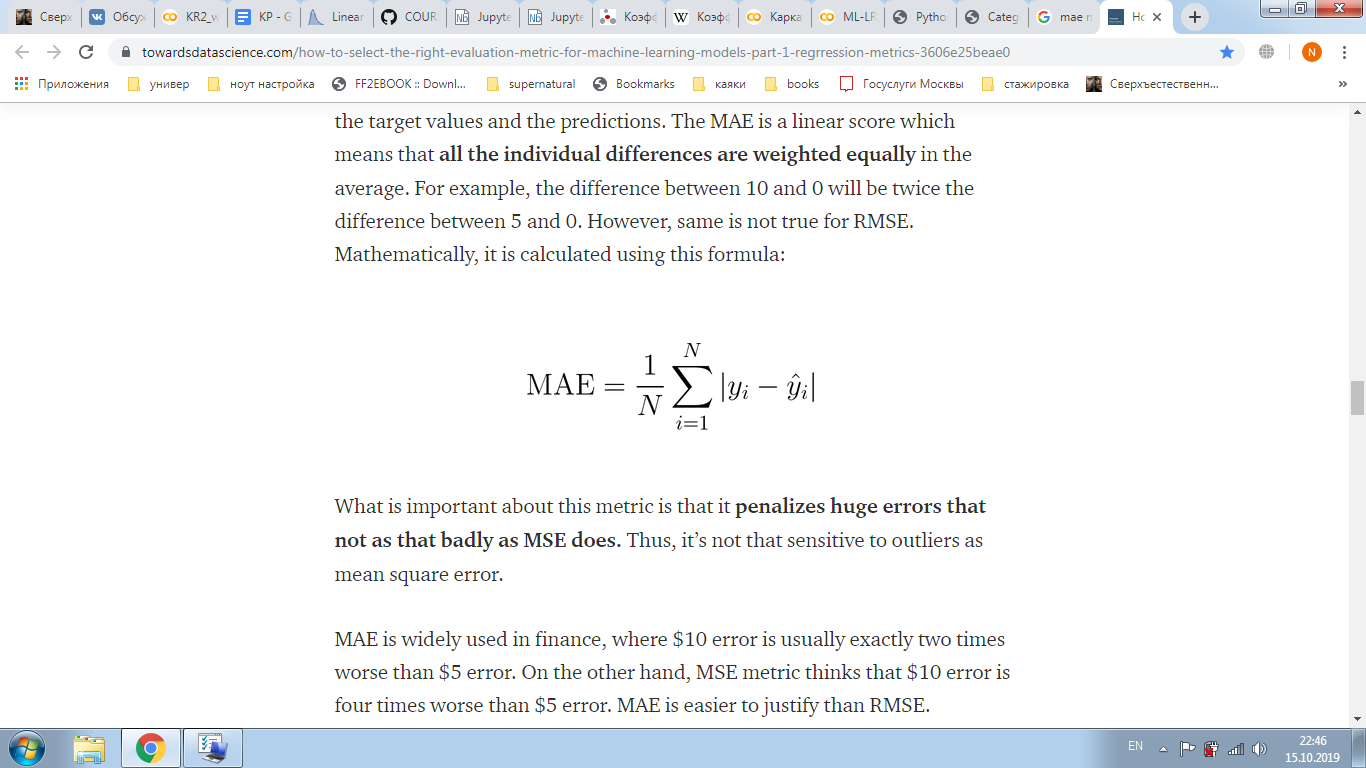


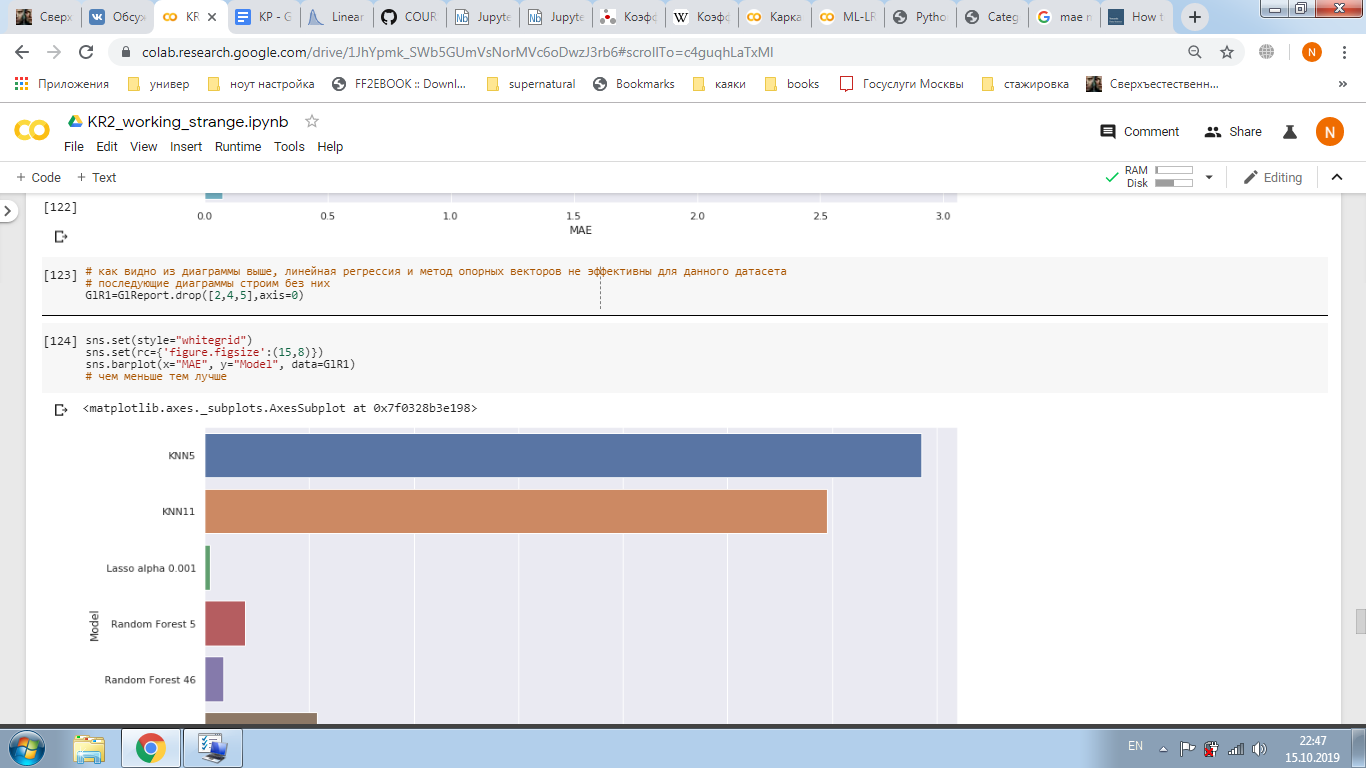


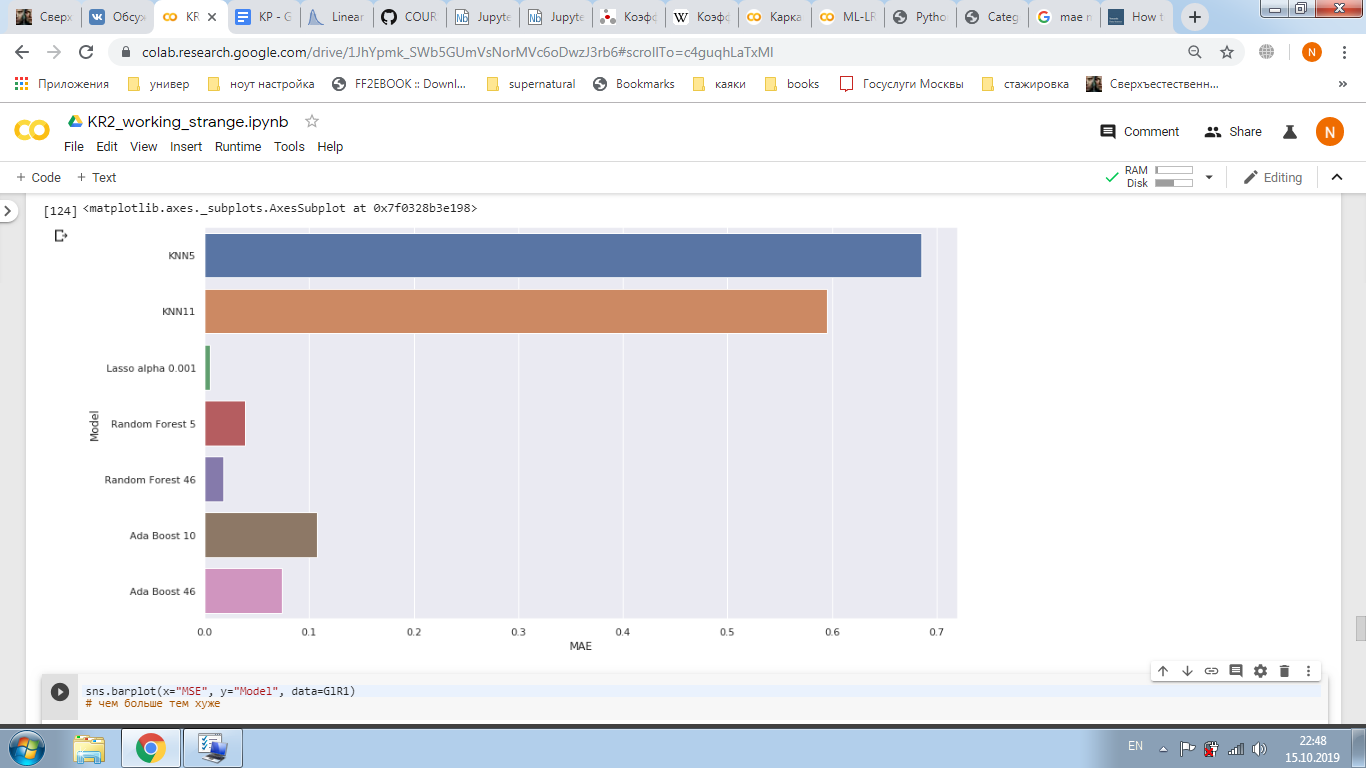


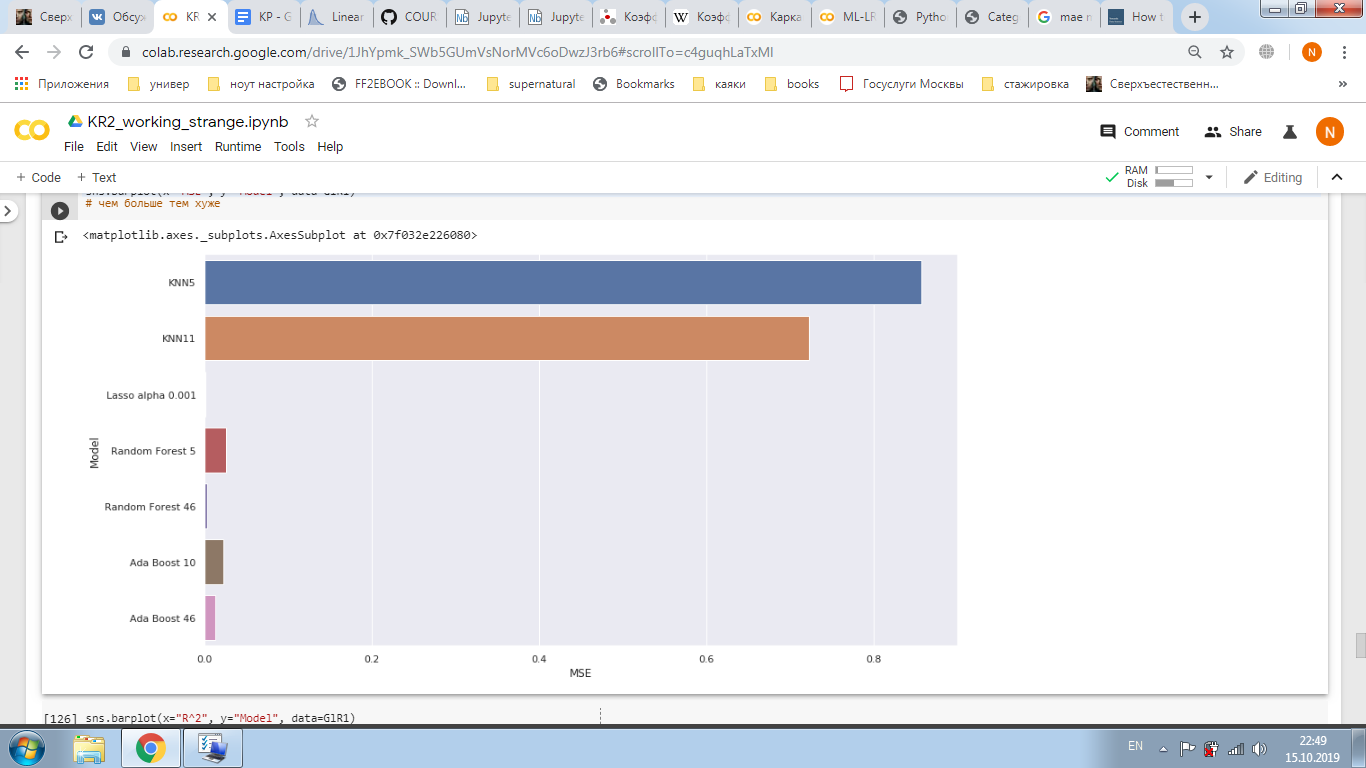


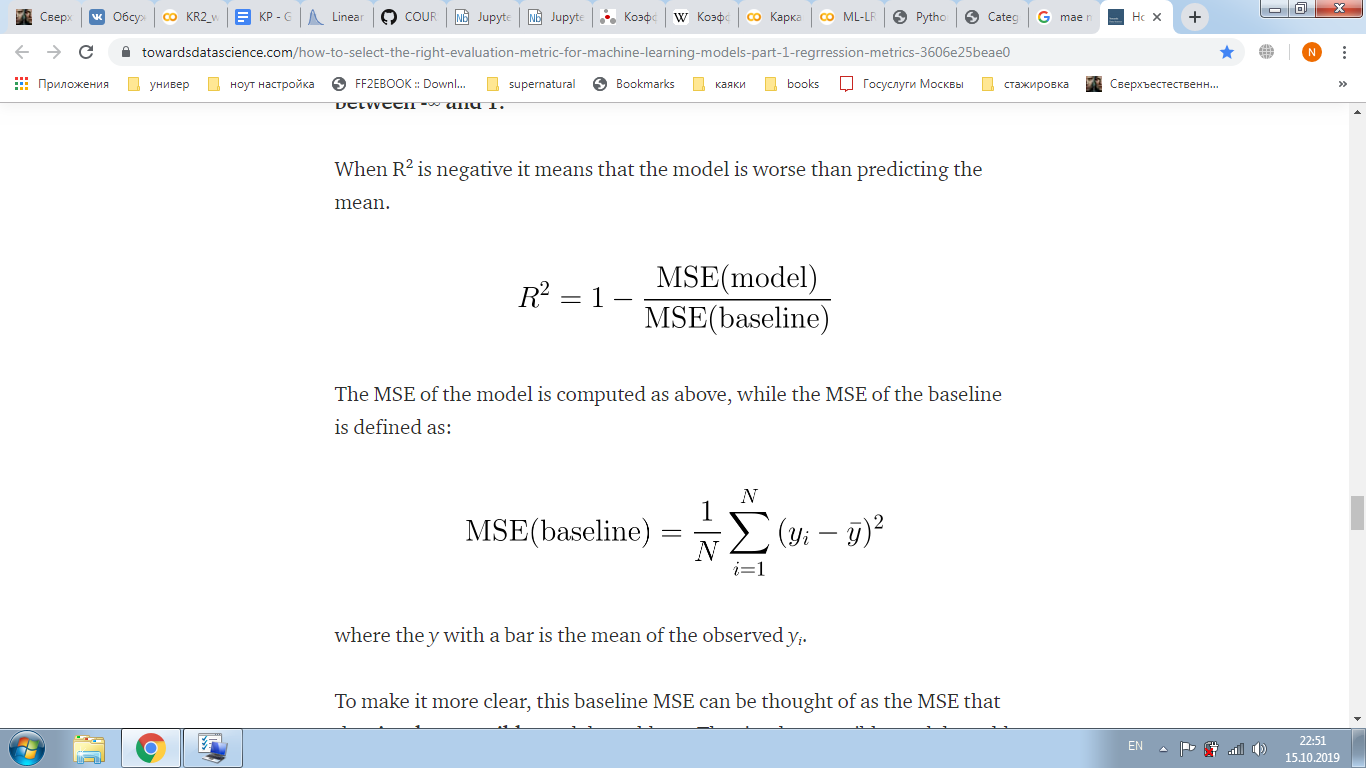
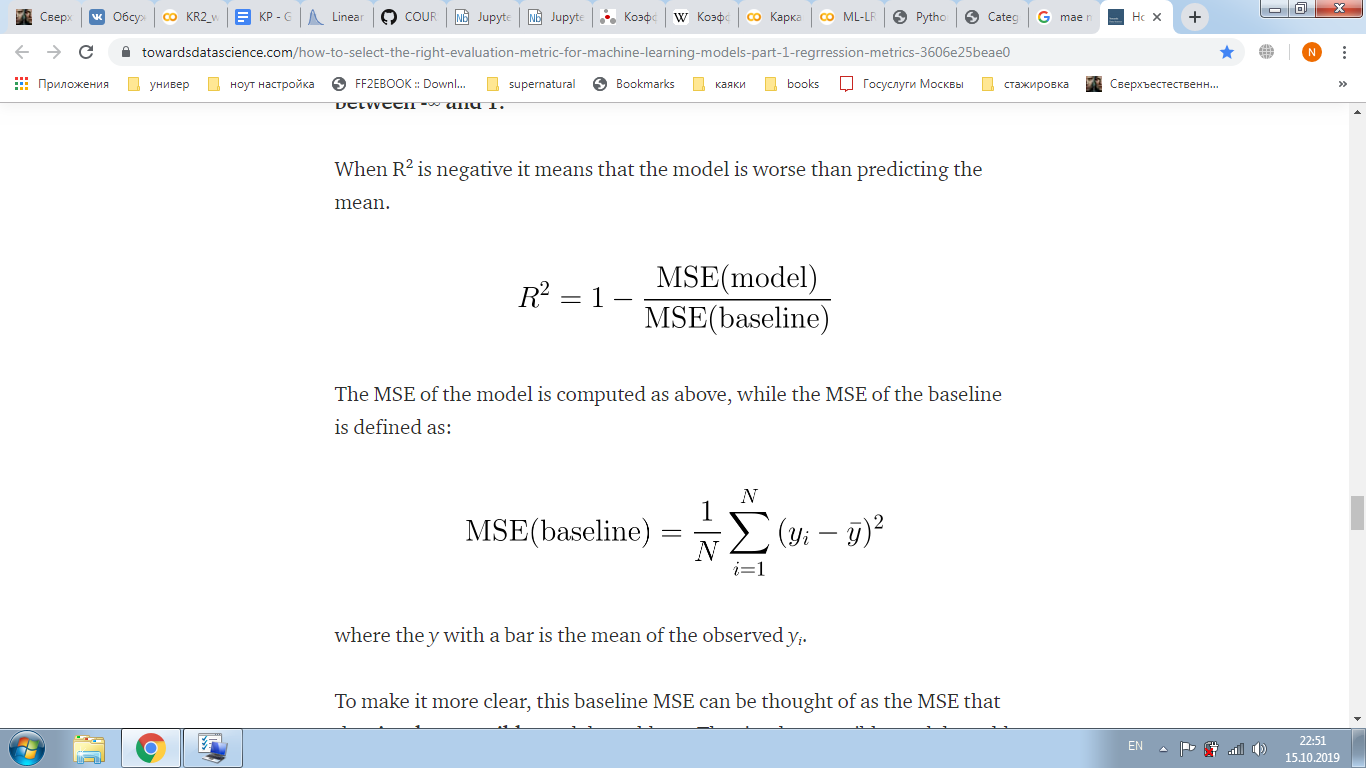
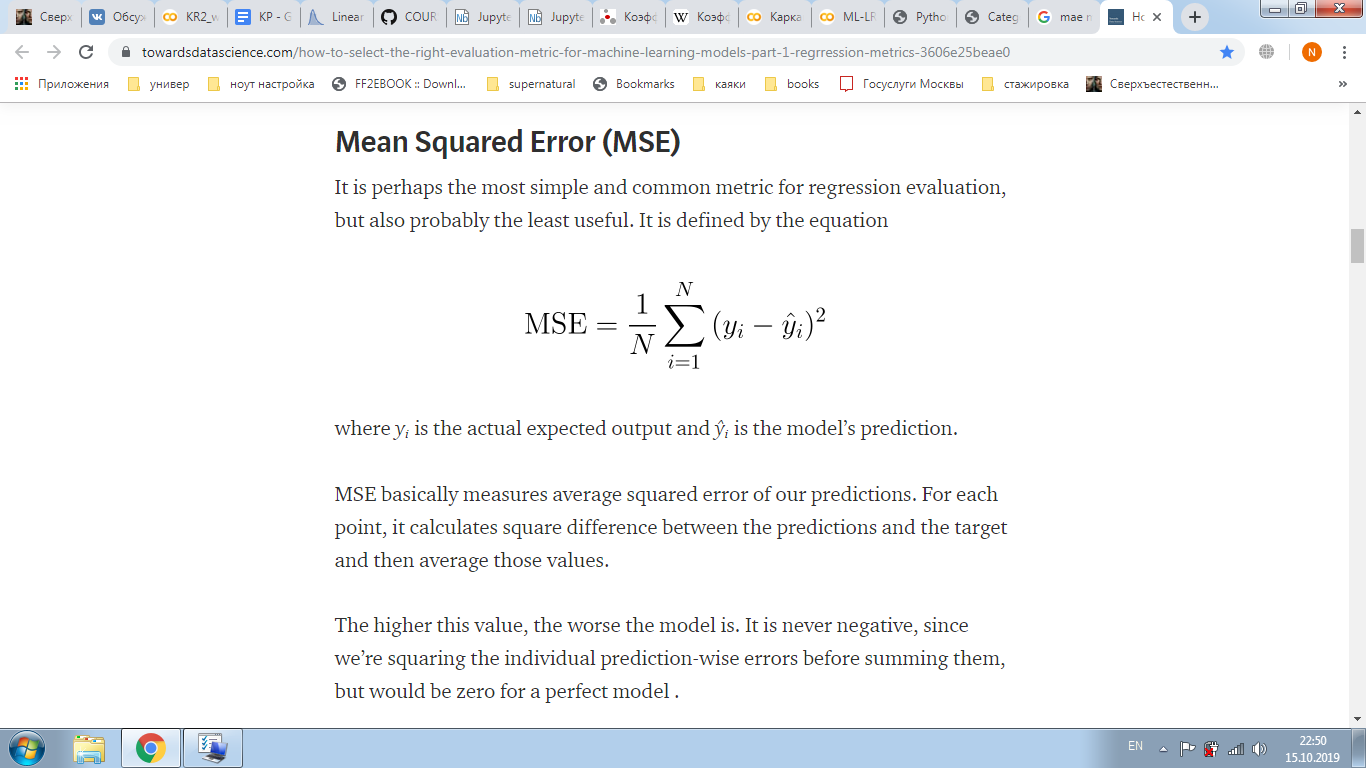


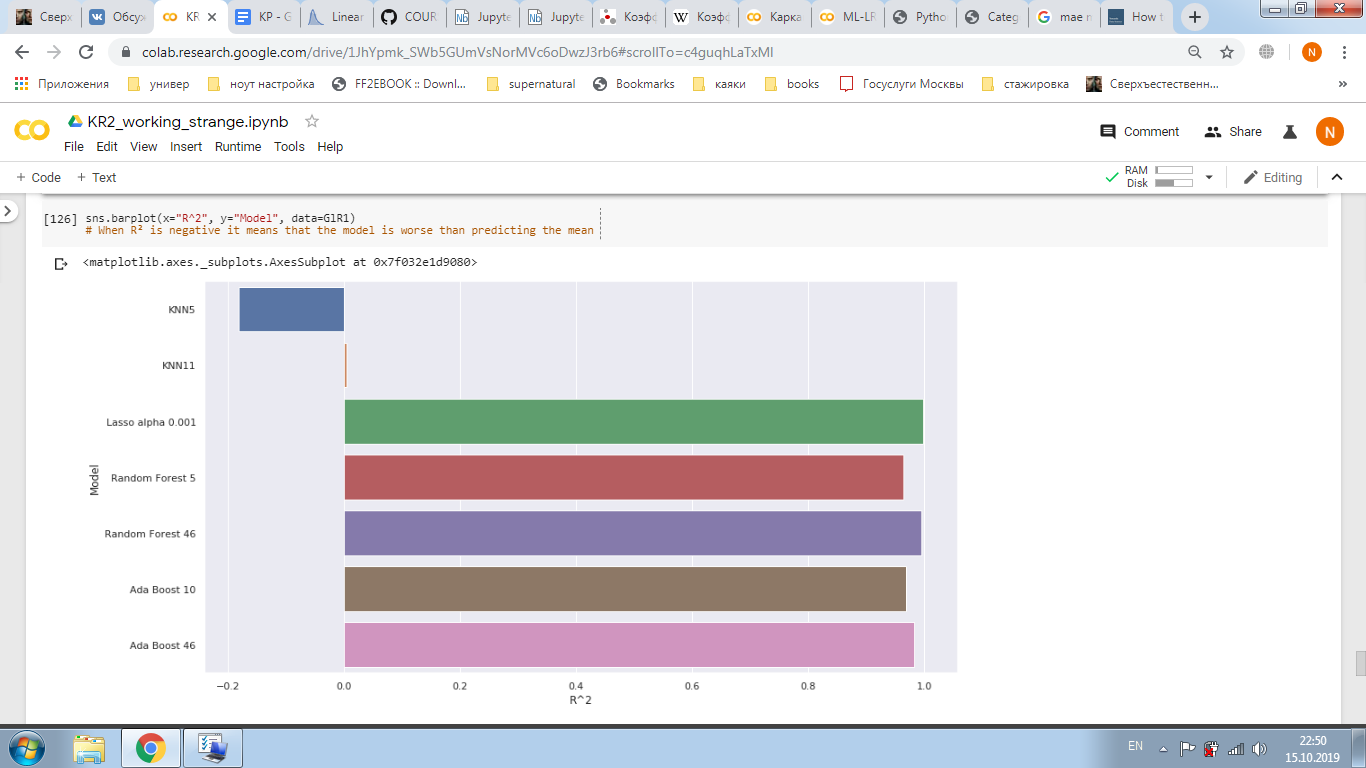


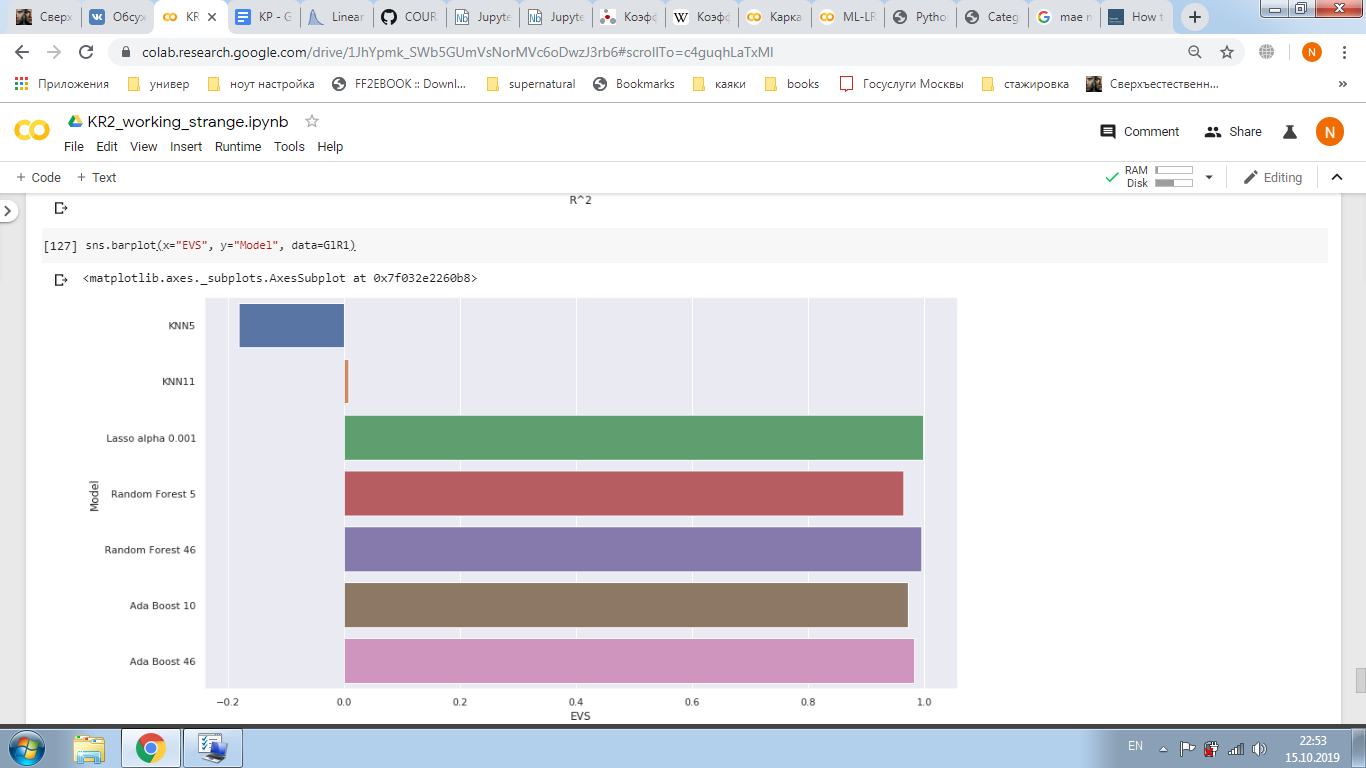












Explained Variance Score = 1 - [Variance(Ypredicted - Yactual) / Variancey\_actual]

Variance(Ypredicted - Yactual) = (Sum of Squared Residuals - **Mean Error**) / n

**Заключение:**

В результате обучения нескольких моделей машинного обучения наиболее удачной для данной задачи оказалась модель линейной регрессии, набрав наилучшие результаты по всем метрикам после получения гиперпараметра через решетчатый поиск с перекрёстной проверкой. Так же хорошие результаты показала модель случайного леса, получение гиперпараметра улучшило ее не значительно. Модель, построенная с помощью метода опорных векторов оказалась неудачной.

**Источники:**

1.<https://blog.usejournal.com/a-quick-introduction-to-k-nearest-neighbors-algorithm-62214cea29c7>

2. <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso>  
3. <https://scikit-learn.org/0.20/modules/generated/sklearn.svm.SVR.html>

4.<https://www.shanelynn.ie/using-pandas-dataframe-creating-editing-viewing-data-in-python/>

5. <https://scikit-learn.org/0.21/modules/generated/sklearn.ensemble.AdaBoostRegressor.html>

6. Introduction to Machine Learning with Python: A Guide for Data Scientists 1st Edition by Andreas C. Müller , Sarah Guido

7.<https://towardsdatascience.com/how-to-select-the-right-evaluation-metric-for-machine-learning-models-part-1-regrression-metrics-3606e25beae0>