МГТУ им. Н. Э. Баумана, кафедра ИУ5

курс “Технология машинного обучения”

Курсовая работа

**«**Проектирование модели машинного обучения**»**

ВЫПОЛНИЛИ:

Матюнин Р.А.

Мелисов А.Е.

Бакланов Н.В.

Ролдугин Е.В.

Жильцов А.А.

Группа: ИУ5-61Б

ПРОВЕРИЛ:

Гапанюк Ю.Е.

Москва 2020

# Содержание

**Оглавление**

[Содержание 2](#_Toc42014720)

[Введение 3](#_Toc42014721)

[Подготовка 4](#_Toc42014722)

[Постановка задачи 6](#_Toc42014723)

[Разведочный анализ 8](#_Toc42014724)

[Основная часть 11](#_Toc42014725)

[Заключение 19](#_Toc42014726)

[Список литературы 20](#_Toc42014727)

# Введение

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

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

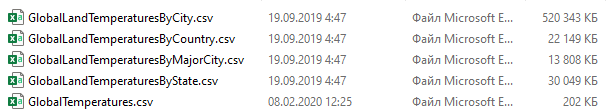
Целью данной работы является, исследование и анализ данных, представленных в текстовых файла (формата csv), для дальнейшего построения модели машинного обучения в прогнозирования климатических изменений.

# Подготовка

В качестве датасета для данной работы, я выбрал данные о климатических изменениях на поверхности Земли, собранные и предоставленные на сайте:

* <https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data/data>

Датасет представляет из себя набор данных мировых температур от 1750 до 2015 гг. и содержит в себе несколько файлов:



GlobalTemperaturesByCity.csv содержит следующие поля:

* **Dt:** поле даты, которая содержит месяц и год.
* **AverageTemperature:** средняя температура.
* **AverageTemperatureUncertainty:** 95% доверительный интервал от среднего значения температуры.
* **City:** город.
* **Country:** страна.
* **Latitude:** долгота.
* **Longitude:** широта.

GlobalTemperaturesByCountry.csv содержит следующие поля:

* **Dt:** поле даты, которая содержит месяц и год.
* **AverageTemperature:** средняя температура.
* **AverageTemperatureUncertainty:** 95% доверительный интервал от среднего значения температуры.
* **Country:** страна.

GlobalTemperaturesByMajorCity.csv содержит следующие поля:

* **Dt:** поле даты, которая содержит месяц и год.
* **AverageTemperature:** средняя температура.
* **AverageTemperatureUncertainty:** 95% доверительный интервал от среднего значения температуры.
* **City:** город.
* **Country:** страна.
* **Latitude:** долгота.
* **Longitude:** широта.

GlobalTemperaturesByState.csv содержит следующие поля:

* **Dt**: поле даты, которая содержит месяц и год.
* **AverageTemperature**: средняя температура.
* **AverageTemperatureUncertainty**: 95% доверительный интервал от среднего значения температуры.
* **State**: область.
* **Country**: страна.

GlobalTemperatures.csv содержит следующие поля:

* **Date**: поле даты, которая содержит месяц и год. Поля 1750-1850 гг. содержат информацию только о средних температурах (**LandAverageTemperature** и **LandAverageTemperatureUncertainty)**, 1850-2015 гг. еще и о максимальной и минимальной температуры земли и мировой температуры океана и земли.
* **LandAverageTemperature**: средняя температура Земли (в Цельсиях).
* **LandAverageTemperatureUncertainty**: 95% доверительный интервал от среднего значения температуры
* **LandMaxTemperature**: максимальная температура Земли (в Цельсиях)
* **LandMaxTemperatureUncertainty**: 95% доверительный интервал от максимального значения температуры
* **LandMinTemperature**: минимальная температура Земли (в Цельсиях)
* **LandMinTemperatureUncertainty**: 95% доверительный интервал от минимального значения температуры
* **LandAndOceanAverageTemperature**: средняя температура суши и океана (в Цельсиях)
* **LandAndOceanAverageTemperatureUncertainty**: 95% доверительный интервал от среднего значения температуры суши и океана/

# 

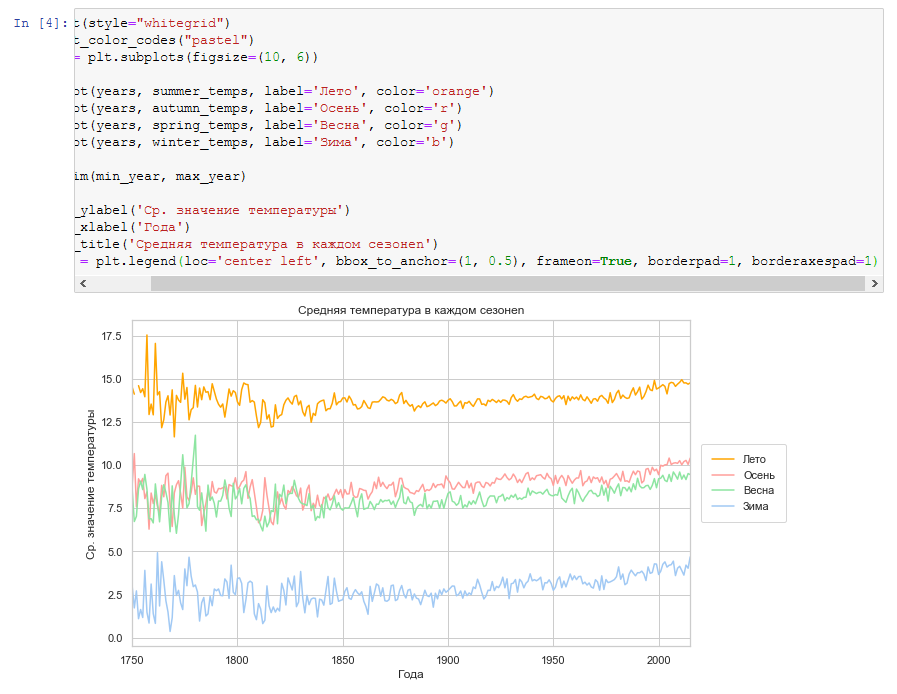
# Постановка задачи

Климатические изменения - всем известный факт. С каждым годом мы замечаем, какие последствия несут за собой эти изменения. С современными технологиями мы можем быть готовыми встретить эти изменения и свести потери к минимуму.

Давайте наглядно рассмотрим данную проблему.







Как видно из графиков (если учесть возможность недостоверности данных 18-19 вв.) можно заметить рост температуры более чем на 2 градуса, особенно зимой.

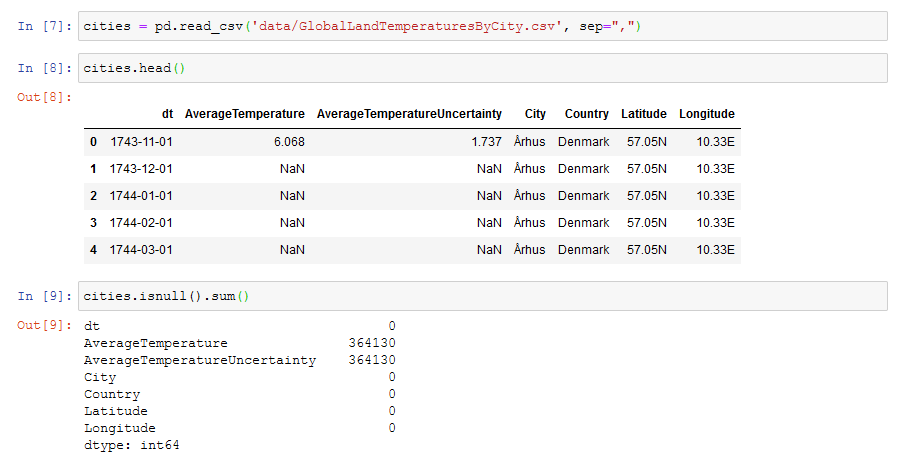
Давайте рассмотри данную проблему более подробнее.

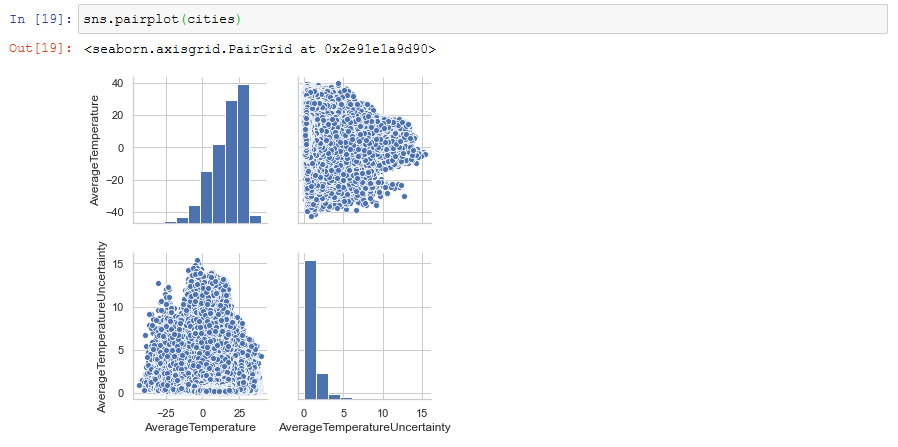
# Разведочный анализ

Подключим нужные библиотеки.



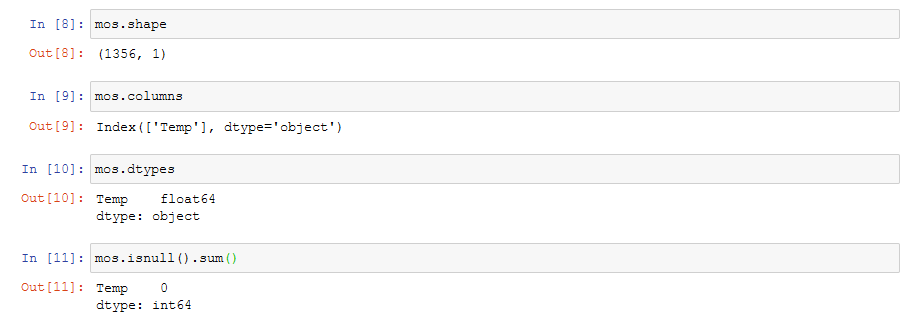
Просмотрим файл GlobalLandTemperaturesByCity.csv:





Изменим наш датасет преобразовав часть файла, связанного только с Калининградом:





Представленный набор данных уже не содержит пропусков

Т.к. представленный датасет весьма специфический, обычные модели для решения задач классификации или регрессии не подойдут.

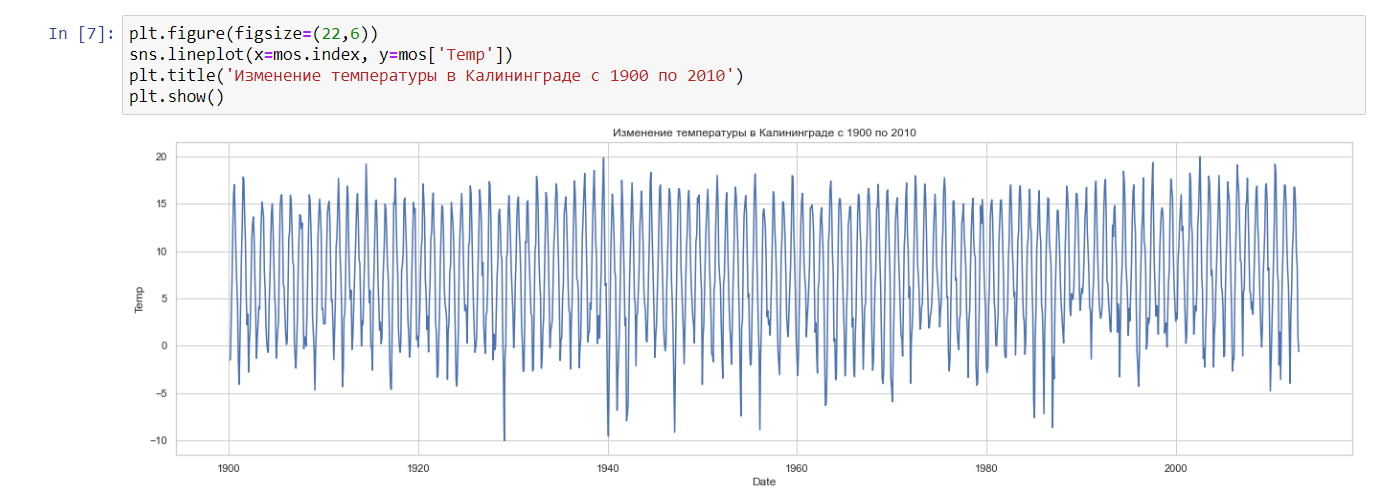
Т.к. существуют некоторые переменные, которые зависят от времени, к примеру, некоторые величины могут иметь эффективное отношение к величинам, которые произошли в прошлом. В настоящее время существует несколько типов моделей прогнозирования временных рядов, поэтому для данной работы я решил использовать сезонные модели ARIMA.

SARIMA расшифровывается как Seasonal Auto Regressive Integrated Moving Average - Сезонная Авто Регрессивная Интегрированная Скользящая Средняя.

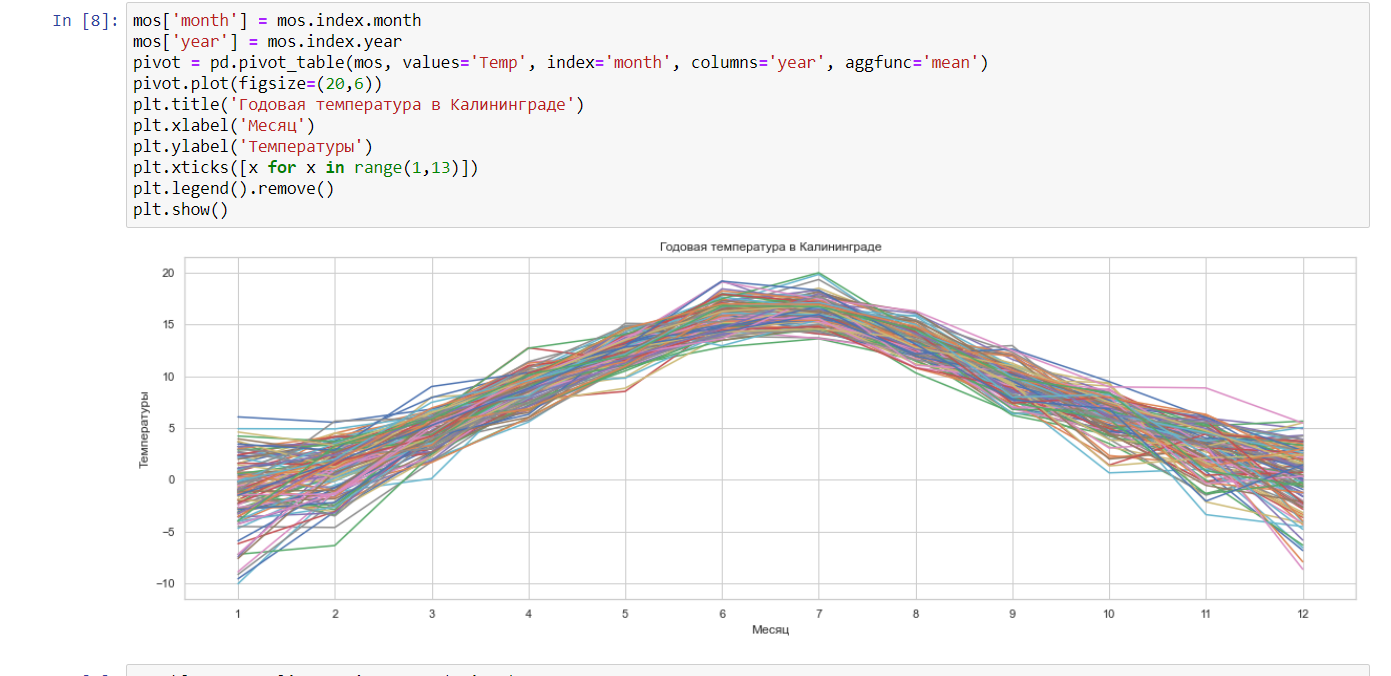
В статистике и эконометрике, в частности в анализе временных рядов, модель авторегрессионного интегрированного скользящего среднего (ARIMA) является обобщением модели авторегрессионного скользящего среднего (ARMA). Обе эти модели адаптированы к данным временного ряда либо для лучшего понимания данных, либо для прогнозирования будущих точек в ряду (прогнозирование).

# Основная часть

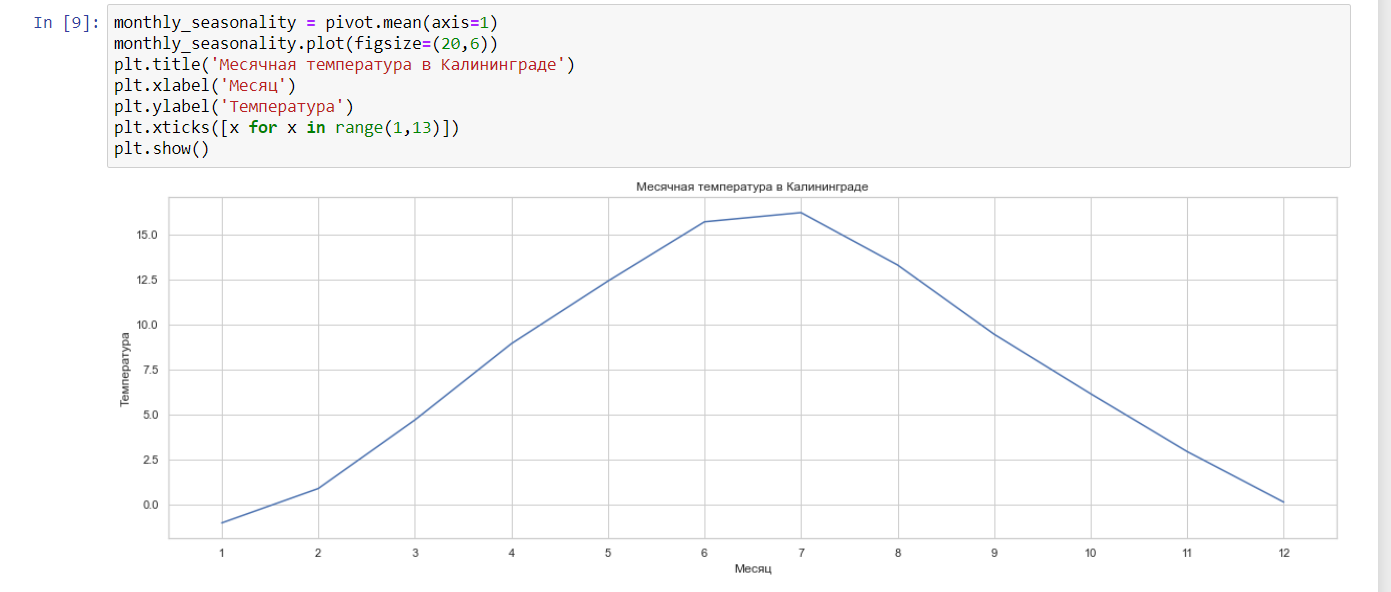
Посмотрим какие изменения температуры произошли в Калининграде с период с 1900 по 2010:



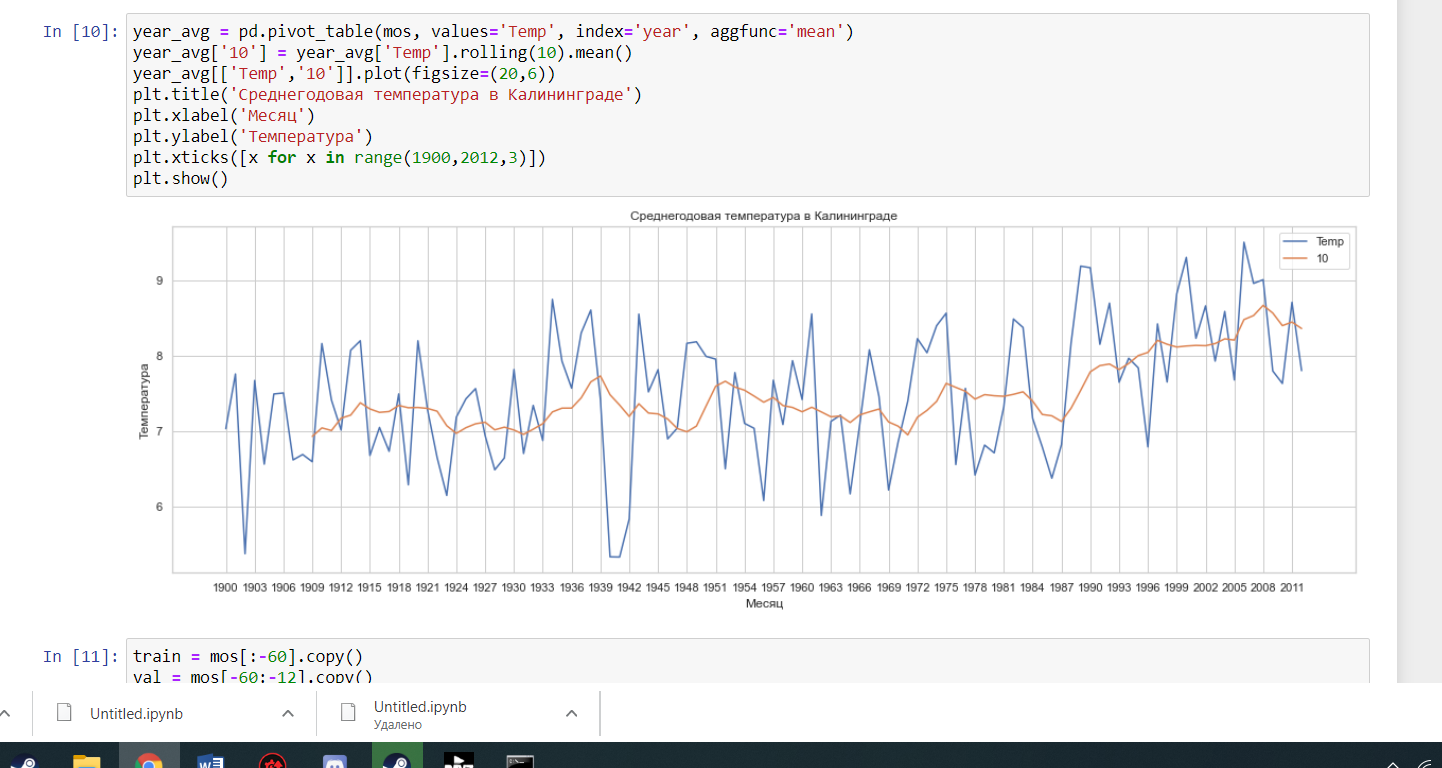
Можно так же заметить, что ряд явно имеет некоторую сезонность: более высокие температуры около июня и июля, и более низкие в декабре и январе.



Давайте объединим все это в один график:



Теперь проверим, существует ли какая-либо тенденция:



И как видно она существует. Средняя температура увеличилась с 7.7° до 8.7°.

Теперь прежде, чем создать прогноз, создадим базовый прогноз в наборе проверки. Модель будет рассматривать предыдущий месяц в качестве базового прогноза на следующий месяц:



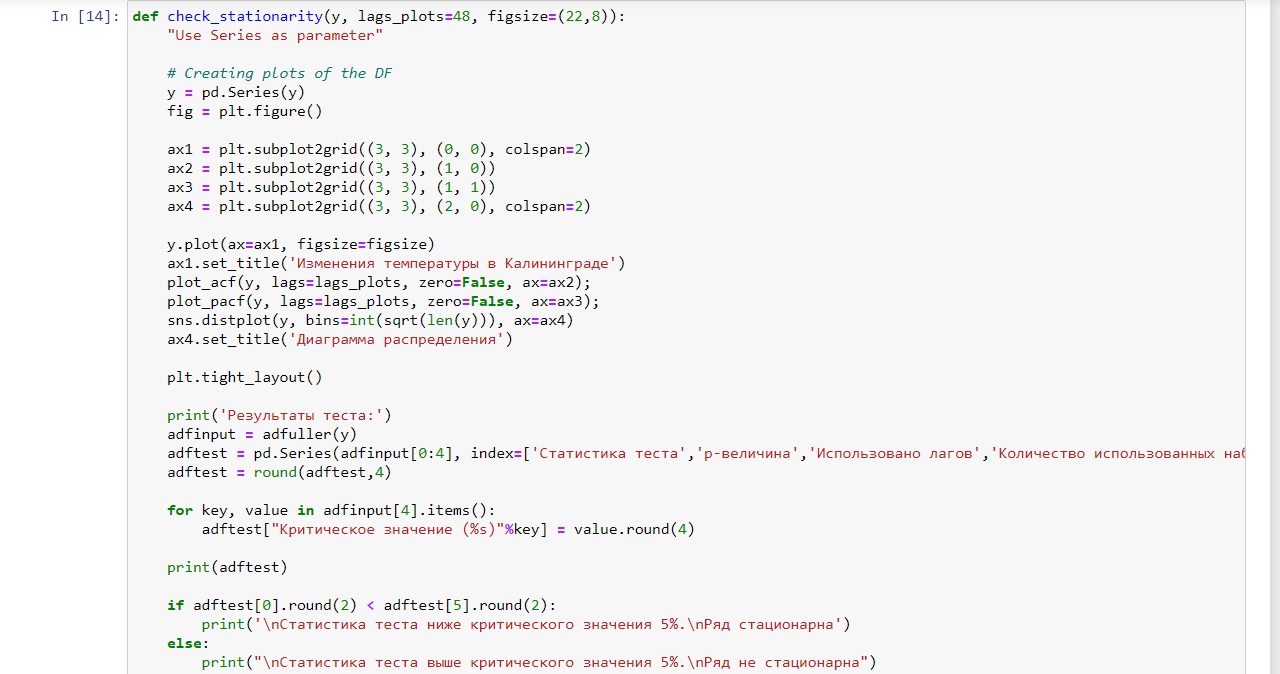
Как мы видим, ряд имеет небольшой восходящую тенденцию, и кажется, что существует некоторая сезонность с более низкими температурами в начале и в конце года и более высокими температурами в середине года.

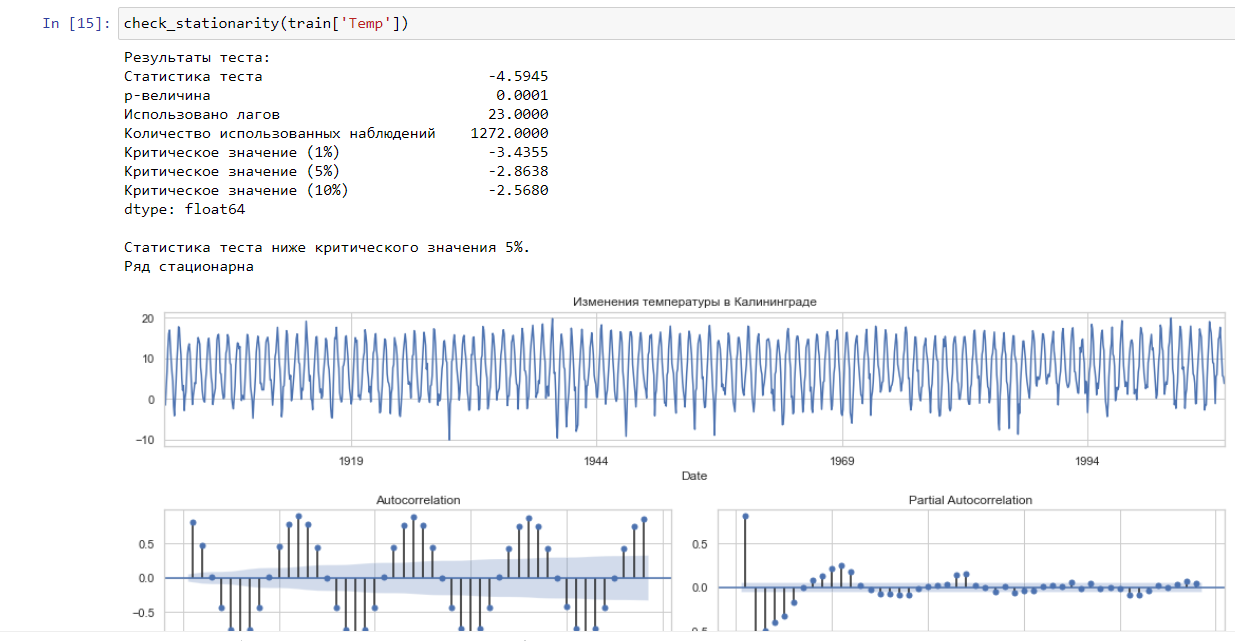
Для создания прогноза временного ряда ряд должен быть стационарным (постоянное среднее, дисперсия и автокорреляция).

Один из способов проверить, является ли ряд стационарным, заключается в использовании функции adfuller. Если значение P меньше 5% (обычное число, используемое для такого рода исследований), серия является стационарной, и вы можете начать создавать свою модель.

Ниже приведена функция проверки стационарности, она показывает:

* Сам ряд.
* Автокорреляционная функция (АКФ):
  + Она показывает корреляцию между текущими температурами и запаздывающими версиями себя.
* Частичная автокорреляция (PACF):
  + Она показывает корреляцию между текущими температурами и лаговой версией, исключая эффекты более ранних лагов, например, он показывает эффективное влияние лага 3 на текущие температуры, исключая эффекты лагов 1 и 2.

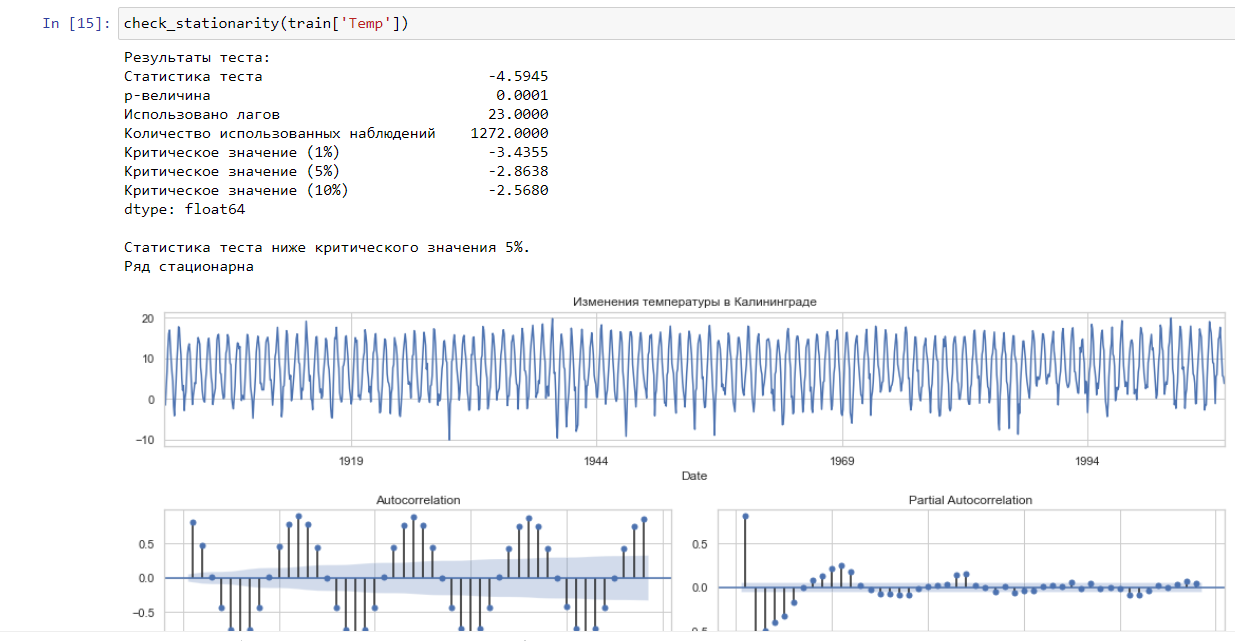




Ряд имеет интересное поведение, есть последовательная значительная отрицательная автокорреляция, повторяющаяся каждые 12 месяцев, это из-за разницы в сезонах, если сегодня зима с холодными температурами через 6 месяцев, у нас будут более высокие температуры летом, поэтому возникает отрицательная автокорреляция. Эти температуры обычно идут в противоположных направлениях.

PACF показывает положительный всплеск в первом запаздывании и снижение к отрицательному PACF в следующих запаздываниях.

Такое поведение между графиками ACF и PACF предполагает модель AR (1), а также первую сезонную разницу. Я снова нанесу на график функцию стационарности с первой сезонной разницей, чтобы узнать, понадобится ли нам какой-нибудь параметр SAR (P) или SMA (Q):



Как показано на графиках выше, первые лаги ACF имеют постепенное затухание, в то время как PACF падает ниже доверительного интервала после третьего лага, это AR-сигнатура с параметром 3.

Чтобы начать прогнозирование набора проверки, я собираюсь создать функцию для использования одношагового прогноза во всем наборе проверки и измерения ошибки:



Всегда важно проверять остатки, я собираюсь создать функцию для построения некоторых важных диаграмм, чтобы помочь нам визуализировать остатки.

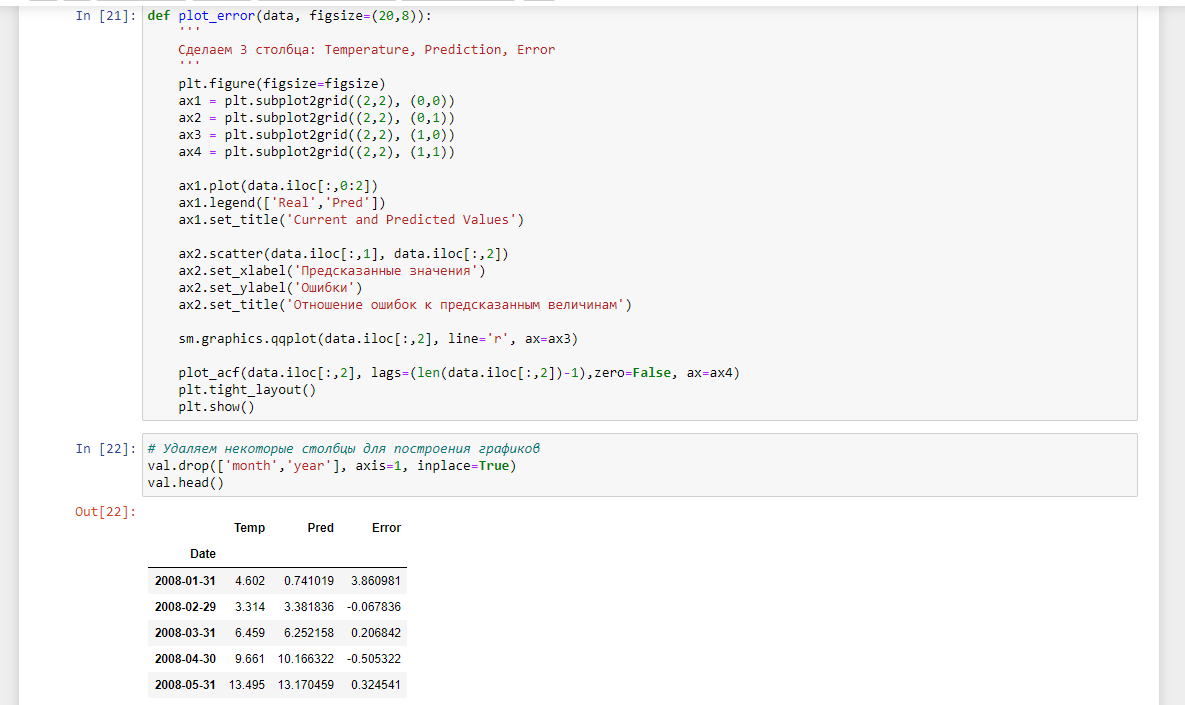
Я собираюсь построить следующие графики:

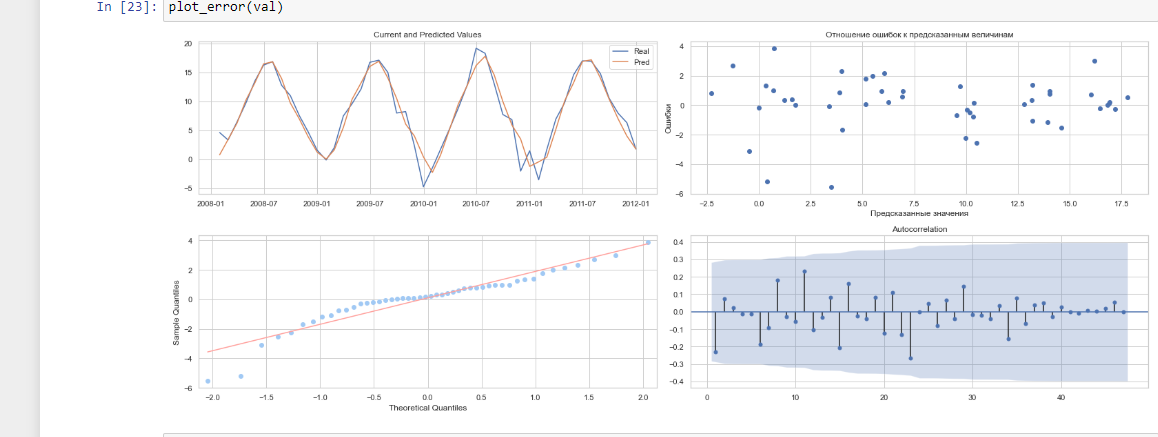
• Текущие и прогнозируемые значения за время.

• Остаточные и прогнозные значения в диаграмме рассеяния.

• График, показывающий распределение ошибок и его идеальное распределение

• График автокорреляции остатков, чтобы увидеть, есть ли еще какая-то корреляция.





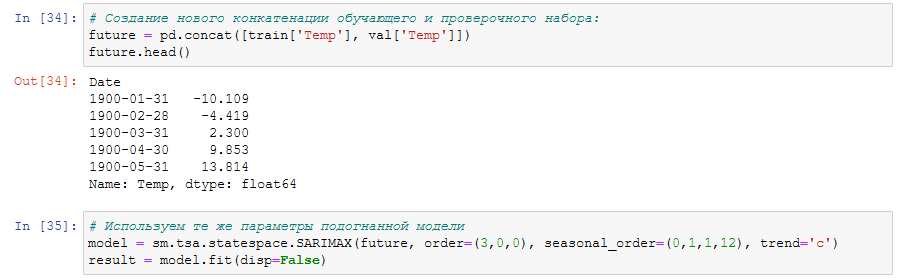
Анализируя вышеприведенные графики, мы видим, что прогнозы очень хорошо соответствуют текущим значениям.

Погрешность в сравнении с прогнозируемыми значениями имеет линейное распределение (погрешности составляют от -8 до +6 при повышении температуры).

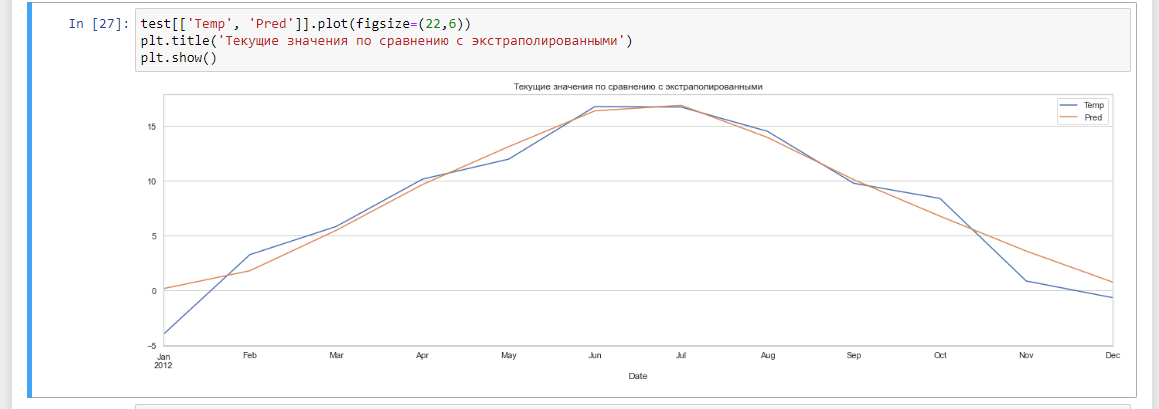
Левый нижний график показывает нормальный паттерн с небольшими выбросами.

График автокорреляции показывает положительный всплеск доверительного интервала чуть выше второго лага, но я считаю, что больше не нужно никаких изменений.

Наконец пришло время экстраполировать прогноз в наборе тестов за последние 12 месяцев:



Создаем новый столбец в тестовом наборе с прогнозируемыми значениями и сравниваем их с реальными значениями.

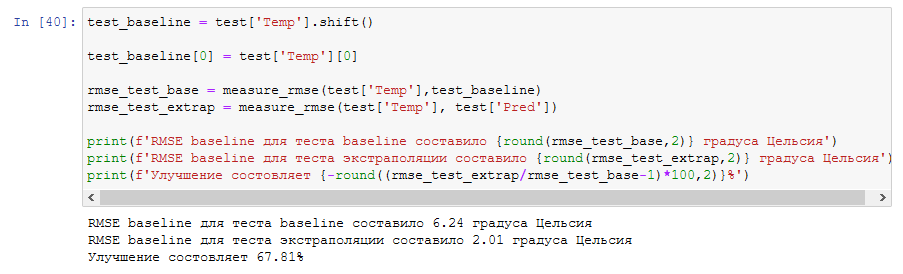


Похоже, что параметры SARIMA были хорошо подогнаны, прогнозируемые значения соответствуют реальным значениям, а также сезонной структуре.

# Заключение

Данная работа показывает, как важен опыт прошлого. Ведь с современными технологиями мы можем предотвратить последствия климатических изменений. Если применить данную модель с более точными данными распределения температуры для других городов России, мы можем получить более точный план примерного изменения температуры. Точное прогнозирование изменения климата поможет людям предвидеть такие катаклизмы, как массовые пожары лесов, наводнение прибрежных к рекам городов и т.д.

Наконец, я оценив модель с RMSE в тестовом наборе (baseline против экстраполяции):



Можно сказать, насколько эффективна данная модель.

# Список литературы

1. Информационные ресурсы по теме SARIMA: <https://machinelearningmastery.com/sarima-for-time-series-forecasting-in-python/>

<https://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average>

1. <https://en.wikipedia.org/wiki/Root-mean-square_deviation>
2. Из курса лекций по основе Анализа данных «Модель авторегрессии и скользящего среднего ARMA(p,q)» Борис Демешев (ВШЭ, Москва): <https://www.youtube.com/watch?v=pQm7ZDgB1tA>
3. «Лекционные материалы по курсу ТМО» Ю.Е. Гапанюк: <https://github.com/ugapanyuk/ml_course_2020/wiki/COURSE_TMO>
4. «Методические указания к курсовой работе» Ю.Е. Гапанюк: <https://nbviewer.jupyter.org/github/ugapanyuk/ml_course_2020/blob/master/common/notebooks/projectexample/project_clas_regr.ipynb>