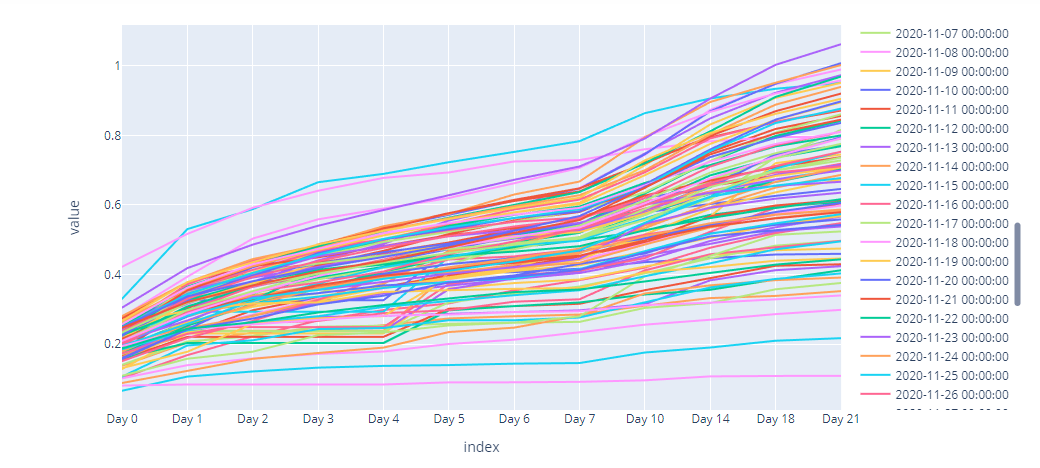
**<https://github.com/smeyanof/data_analysis>**

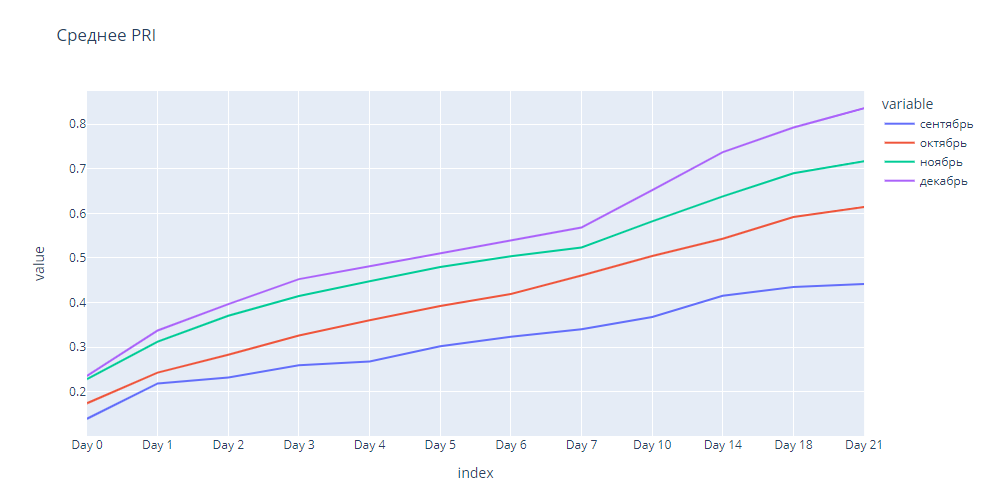
**Summary**

## Построить усредненную кривую RPI на интервале Day0 - Day21 с интерполяцией данных внутри отсутствующих дней.

Посмотрим на данные.

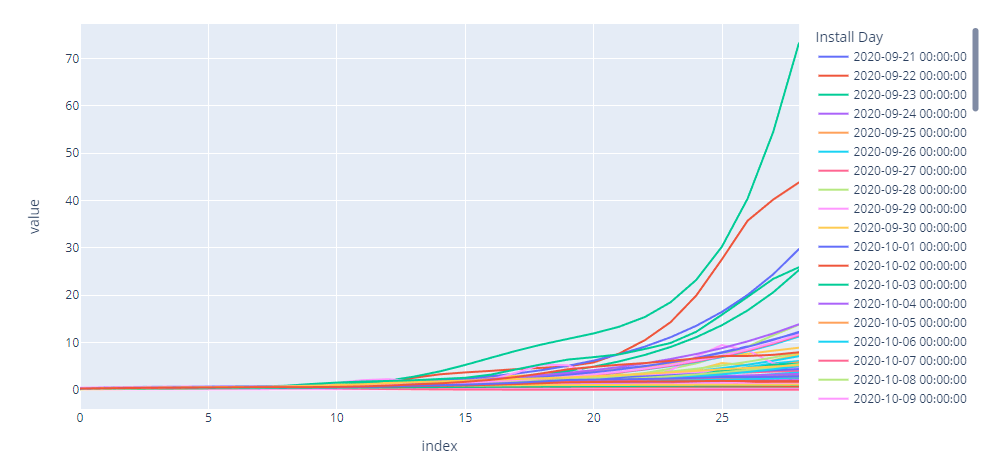
**Графики PRI**

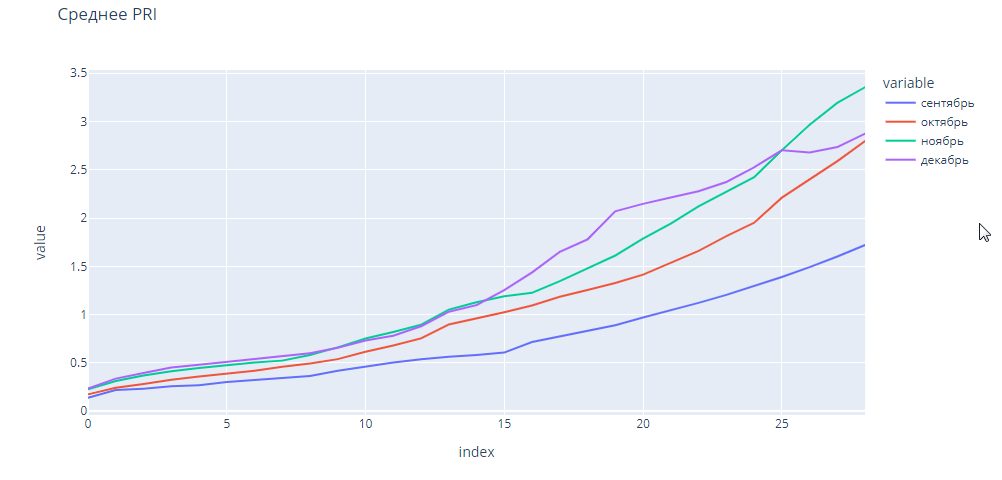




Сначала хотел использовать модель ARIMA для экстраполяции временных рядов. Однако ряды нестационарные, в общем, это не сработало.

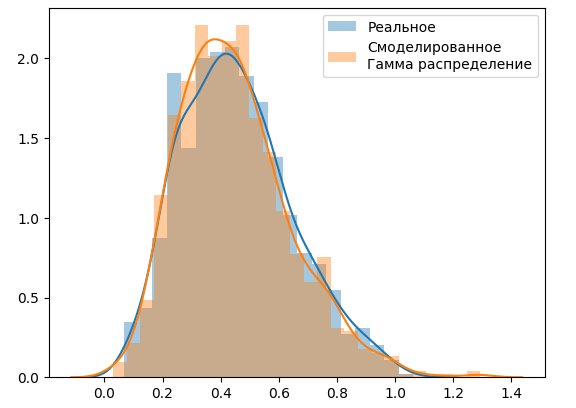
**Спрогнозированные ARIMA значения**



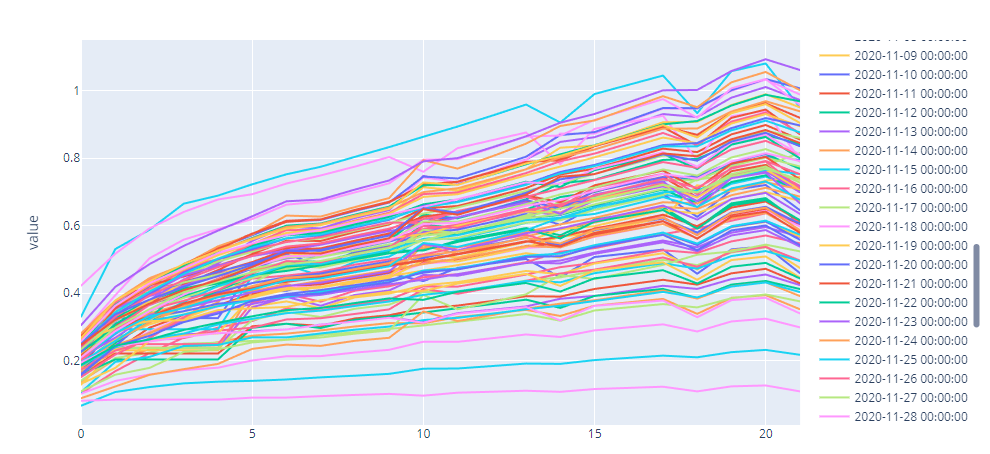


Как видно на графике, среднее завышается более, чем в 3 раза.

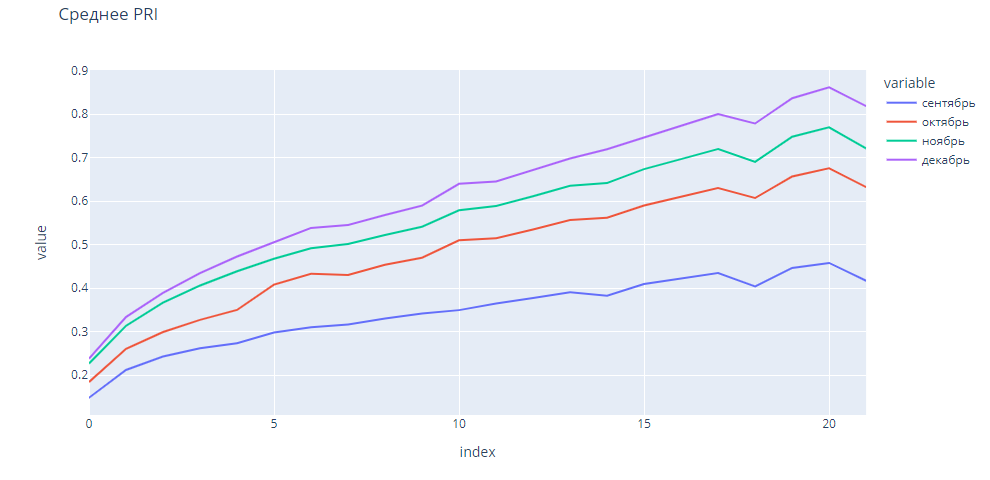
Решил посмотреть распределение, возможно, подходит Гамма.



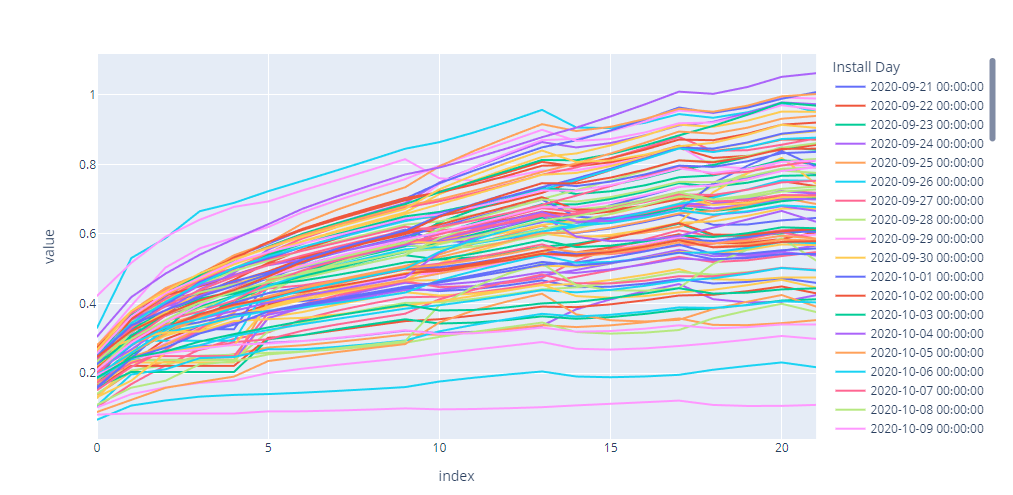
Моделировал данные AR моделью автокорреляции с Гамма распределением.



Провалы в графиках – это реальные данные на 10, 14, 18 и 21 дни. Как видно, модель завышает значения.

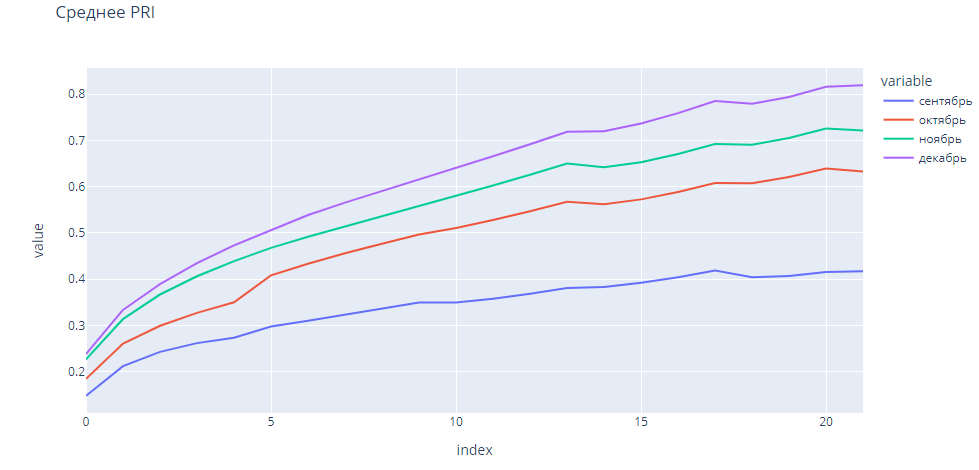


Наконец, взял обычную AR модель с обычной линейно регрессией.

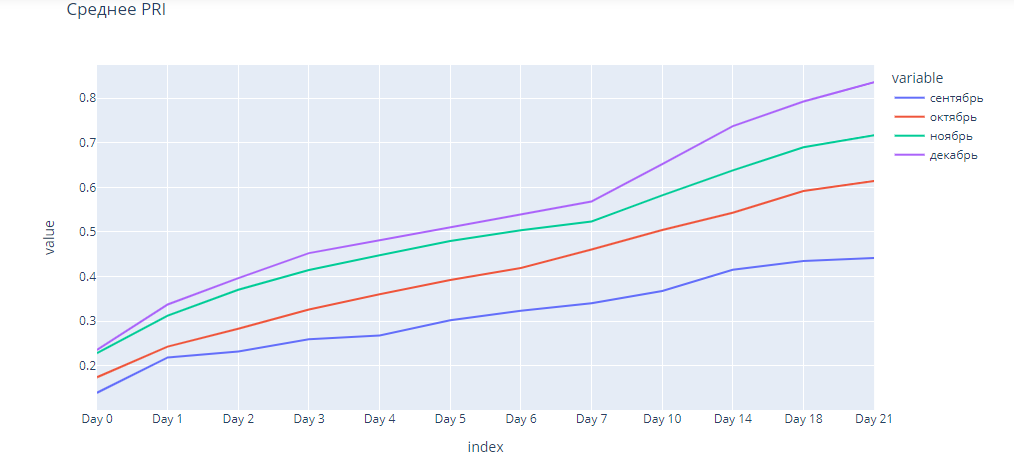


Выглядит неплохо. Тренды сохраняются, значения линейно возрастают.

Посмотрим на среднее. (Это ответ)



И для сравнения на среднее по имеющимся данным.



Как видно, модель хорошо моделирует данные.

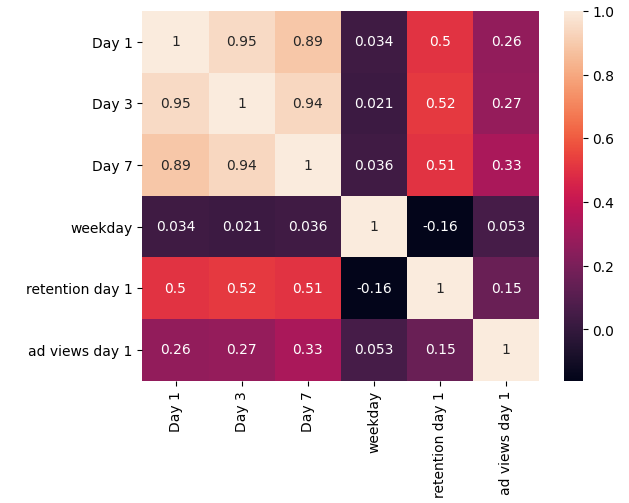
## Выяснить характер зависимости метрики RPI в 1, 3 и 7 дни после установки игры от:

## Дня недели (возможно доход от пользователей, установивших игру в выходные больше, чем от тех, кто установил в будни?)

## Retention - Day 1

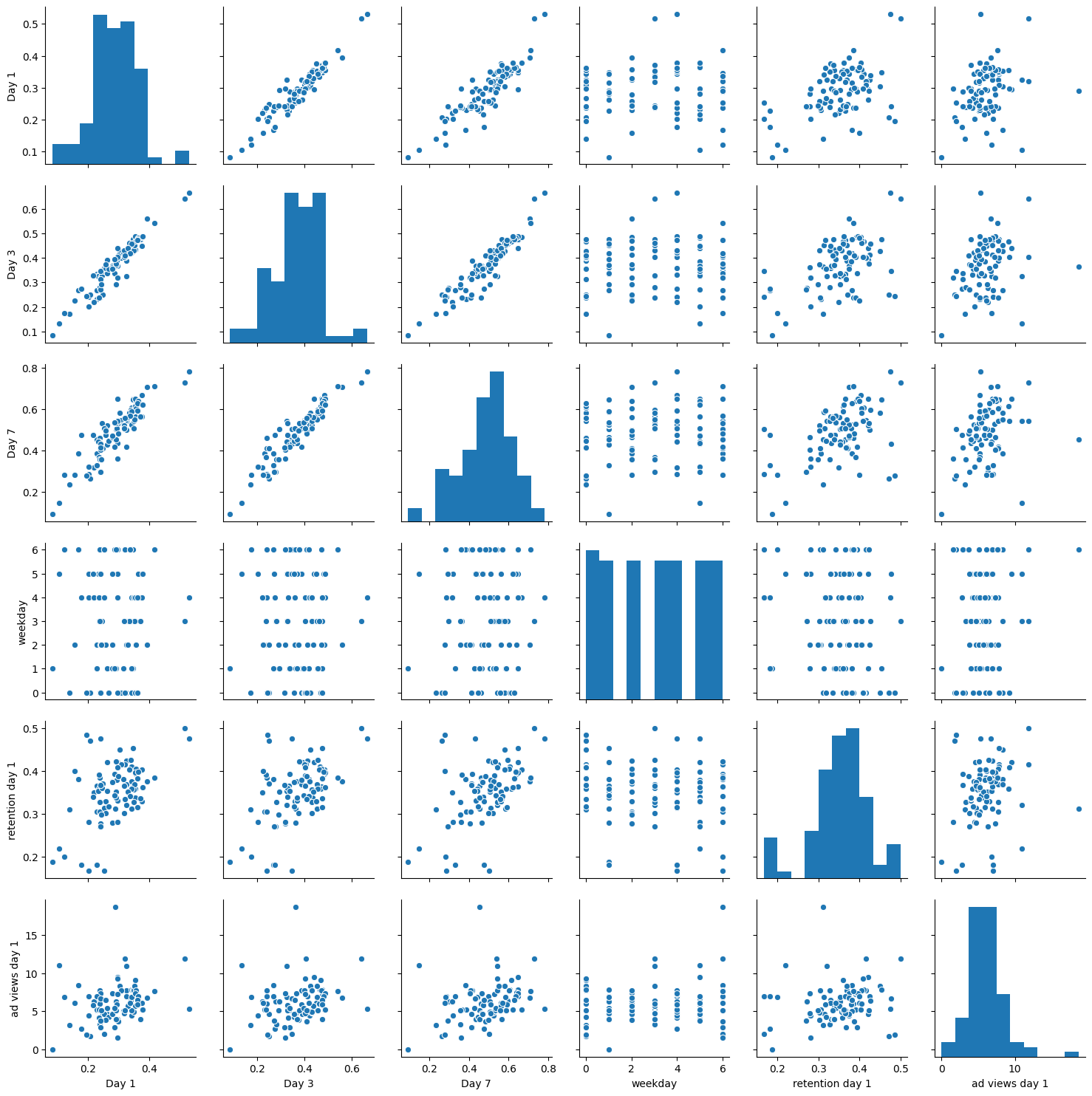
## Ad Views per Day - Day 1

Посмотрел коэффициенты линейно корреляции. (они изменяются от -1 до 1, где 0 – нет зависимости, -1 – обратная линейная зависимость, 1 – линейная зависимость)



Присутствует автокорреляция (конечно). Доход плохо зависит от дня недели установки. А вот Retention и Ad views больше.

Взглянем на парные графики.



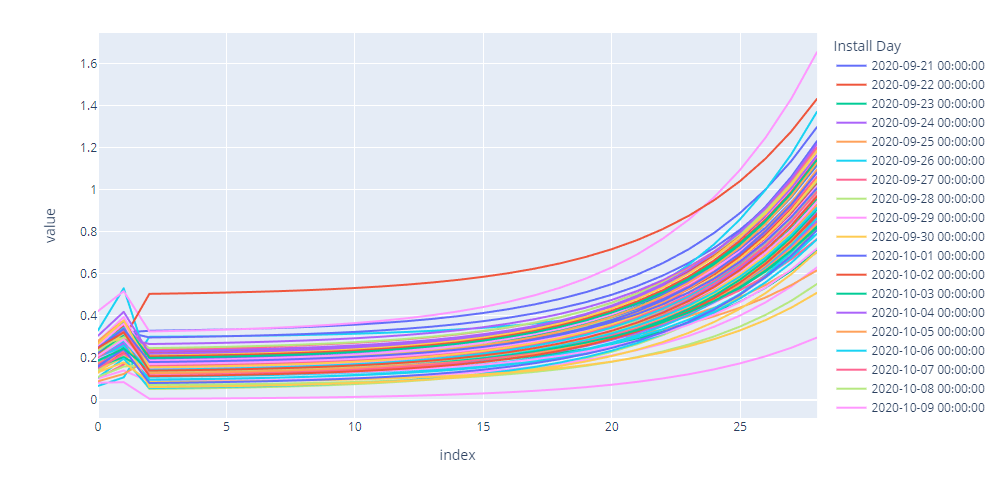
Что и следовало ожидать, между day1, 3, 7 линейная зависимость. Weekday - случайный. Retention и Ad views – можно провести линию.

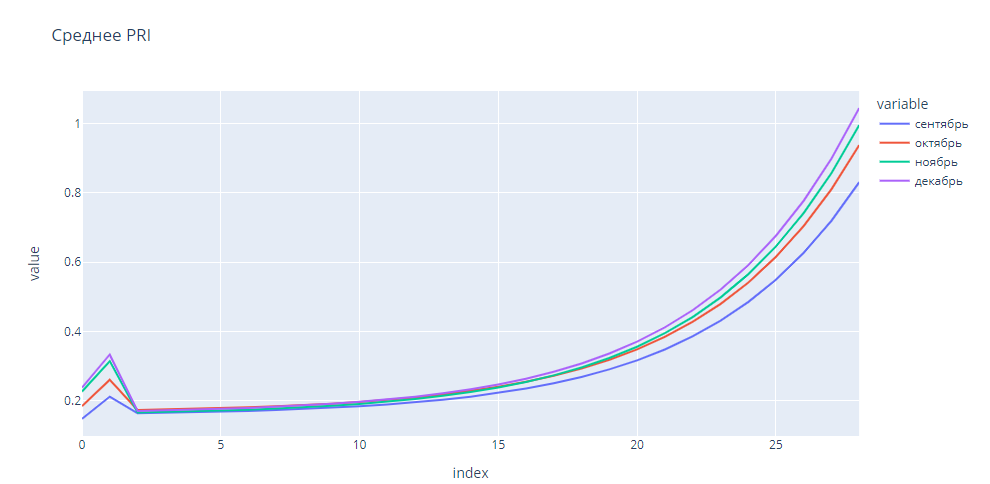
## Построить модель предсказания дохода в когорте на Day28 по значению метрики RPI, Retention, Ad Views per Day на Day1 с минимальной среднеквадратической ошибкой.

Конечно, построение модели на 28 день по известным значениям только на первый день – это колдунство чистой воды.

Поэтому, вспоминаем, что распределение PRI похоже на гамма. Тогда возьмем AR модель c гамма распределением. Подбирать значение alpha для модели буду по значению среднеквадратичной ошибки на 21 день (это значение известно).

Вот что получилось.





Возможно, это могут предсказывать какие-то сложные нейросети.