Диплом

на тему "Restaurant Visitors"

("Посетители ресторана")

Выполнила

Губанова А.В.

2022 г.

Введение

Выбранный датасет относится к ресторанному бизнесу и содержит данные о посещаемости ресторанов. Анализ данной темы интересен для бизнеса, так как позволяет решить следующие задачи:

- Оценка количества посещаемости каждого ресторана, можно сделать вывод какой ресторан посещают чаще, какой реже;

- Анализ взаимосвязи посещаемости каждого ресторана от дня недели, можно сделать вывод в какие дни недели рестораны посещают чаще, в какие реже.

Возможность прогнозирования позволяет планировать развитие на будущее или наоборот отслеживать негативные тренды.

Цель

Проведение исследования данных о посещаемости ресторанов, выявление зависимости посещаемости от дней недели и построение прогноза посещаемости на несколько дней.

Постановка задачи

Основные задачи:

- задача 1. подтвердить зависимость посещаемости от дня недели;
- задача 2. получить прогноз на 2 недели вперед.

Дополнительные:

Обработать данные и отфильтровать необходимую информацию, определить основные метрики, произвести расчет основных стат. показателей, проверить возможные кореляции метрик/атрибутов, проверить выбранные модели на предмет возможности и качества прогнозирования, получить прогнозы и сделать выводы.

Анализ датасета

Перед тем как приступить к анализу данных предварительно изучаем сам датасет на предмет содержащихся данных.

Полученный датасет "Restaurant Visitors" содержит данные о четырех ресторанах одной сети:

- date (дата);
- weekday (день недели);
- holiday (праздник);
- holiday_name (название праздника);
- rest1- rest4 (количество посетителей по ресторанам);
- total (суммарное количество посетителей).

Приводим в порядок типы данных наших атрибутов, проверяем наличие пустых значений, заменяем пустые значения на нули, проверяем наличие отсутствующих дат в интервале наблюдений, сортируем по датам, для исключения несоответствия распределения дат в датасете. После всех внесенных правок оставляем только необходимые для анализа атрибуты: date, weekday, rest1- rest4, total.

Для начала используем разведочный анализ данных для получения статистических значений по каждому атрибуту (таблица 1).

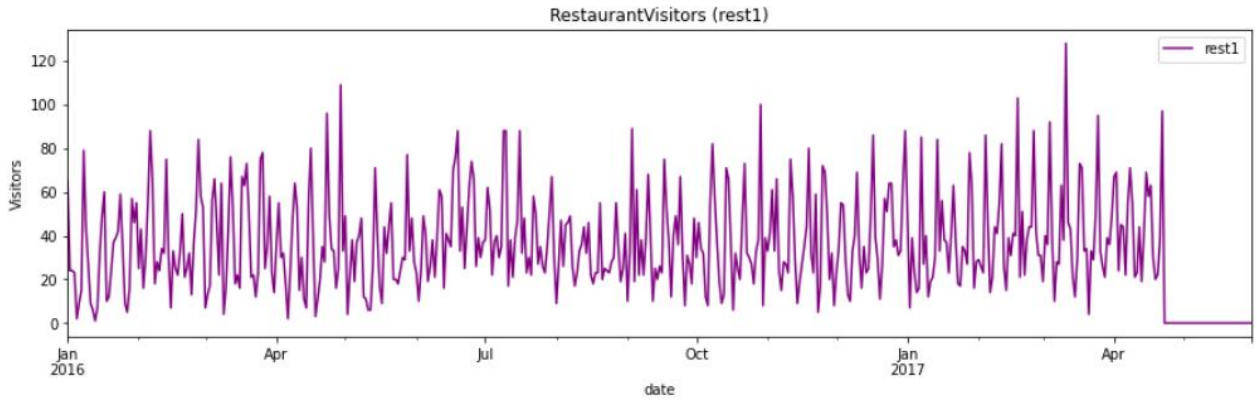
Таблица 1 - Статистические значения по атрибутам

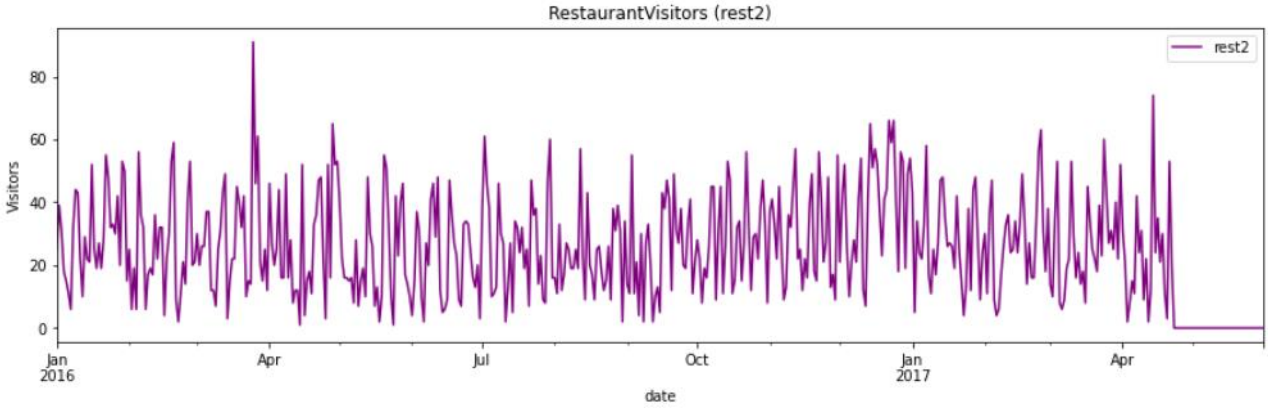
	Rest1	**Rest2**	**Rest3**	**Rest4**	**total**
count	517.00	517.00	517.00	517.00	517.00
mean	34.70	25.11	29.17	34.72	123.70
std	22.93	16.74	19.06	25.66	67.45
min	0.00	0.00	0.00	0.00	0.00
25%	20.00	12.00	17.00	16.00	77.00
50%	32.00	23.00	23.00	28.00	111.00
75%	48.00	37.00	43.00	49.00	171.00
max	128.00	91.00	99.00	139.00	316.00

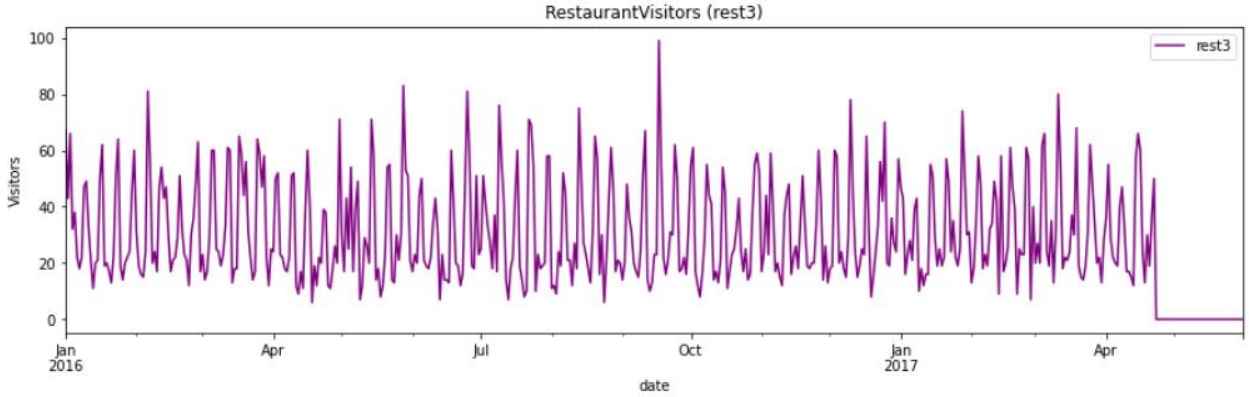
Из таблицы 1 видно, что максимальная посещаемость у четвертого ресторана - 139, далее у первого ресторана -128, у третьего и второго ресторанов 99 и 91 соответственно. Если рассматривать средний показатель посещаемости по ресторанам (mean), то он находится в диапазоне от 25 до 38. Разброс небольшой и можно сделать вывод, что в среднем каждый ресторан в день посещают приблизительно одинаковое количество раз.

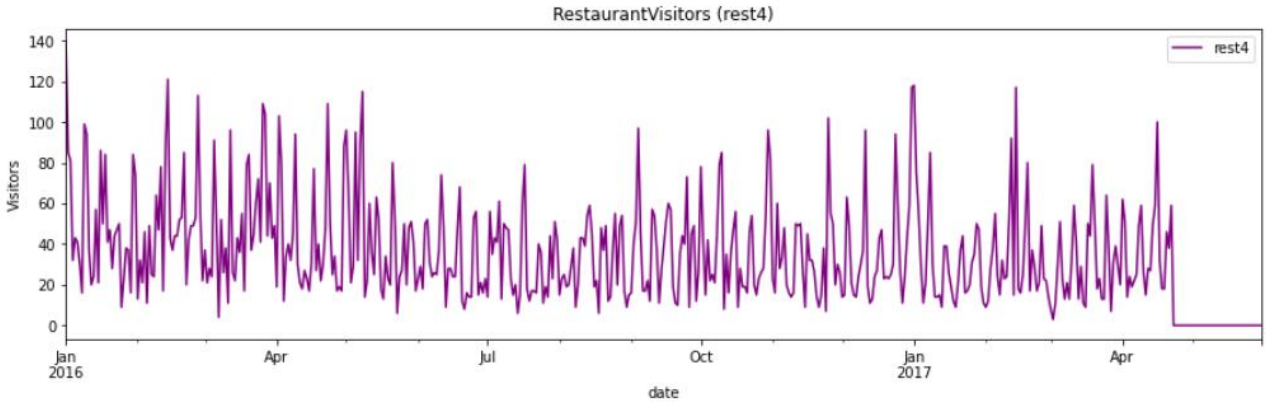
За рассматриваемый период 2016.01 - 2017.05 четвертый ресторан посещали 17948 раз, первый ресторан 17941 раз, третий ресторан -15080 и второй ресторан -12984 раз.

Далее строим графики для отображения каждого атрибута (рисунок 1).


RestaurantVisitors (rest1)
rest1
Visitors
0
20
40
60
80
100
120
Jan
2016
Apr
Jul
Oct
Jan
2017
Apr
date


RestaurantVisitors (rest2)
rest2
Visitors
0
20
40
60
80
Jan
2016
Apr
Jul
Oct
Jan
2017
Apr
date


RestaurantVisitors (rest3)
rest3
Visitors
0
20
40
60
80
100
Jan
2016
Apr
Jul
Oct
Jan
2017
Apr
date


RestaurantVisitors (rest4)
rest4
Visitors
0
20
40
60
80
100
120
140
Jan
2016
Apr
Jul
Oct
Jan
2017
Apr
date

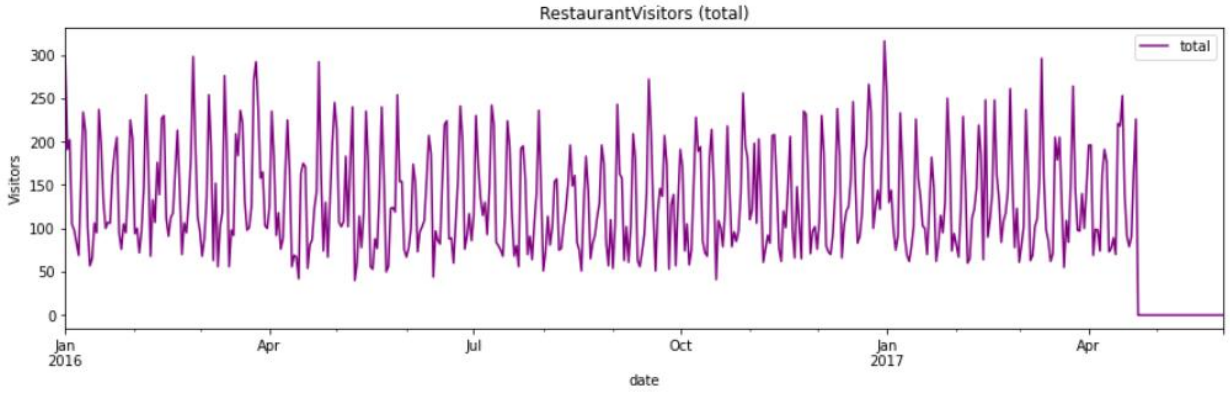
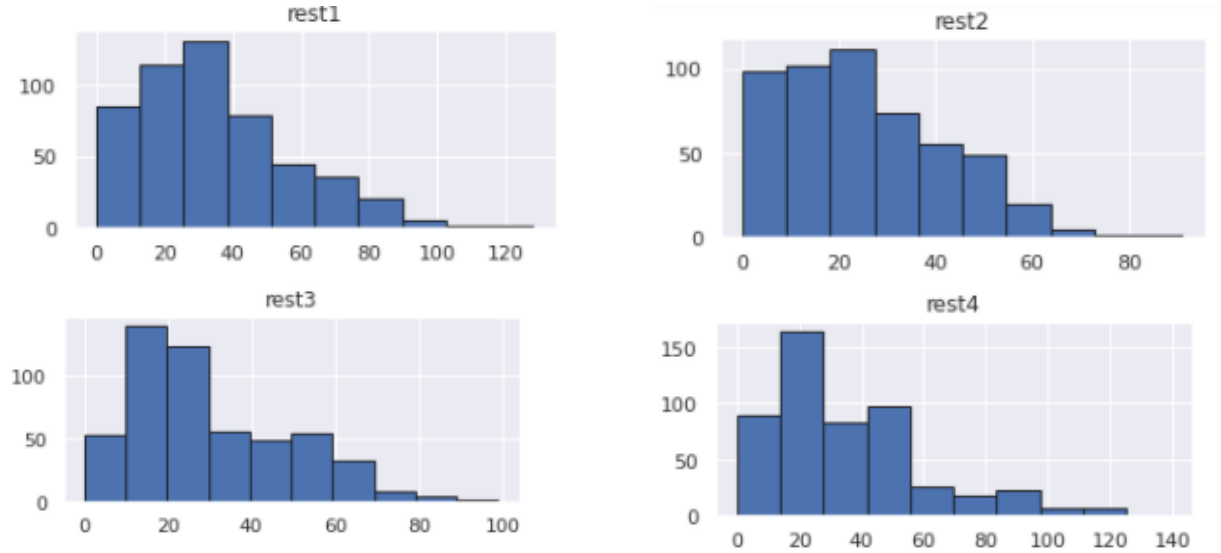


Рисунок 1 – Зависимость посещаемости ресторанов от дней

По графикам видно, что наблюдается недельная сезонность, предположительно связанная с более высокой посещаемостью в выходные дни.

По графикам также четко отслеживается ситуация с пропусками данных, замененных ранее на нули. Данные пропуски располагаются в конце датасета и по факту не связаны с наблюдаемыми значениями. Поэтому в дальнейшем на этапе прогнозирования необходимо исключить данный "нулевой хвост" для более корректного результата по моделям.

Далее для анализа построим гистограммы для определения распределения данных (рисунок 2).



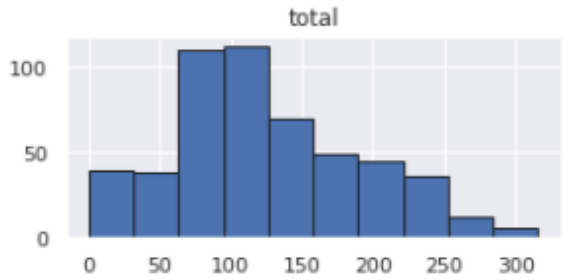


Рисунок 2 – Зависимость посещаемости ресторанов от количества наблюдений

По графикам видно, что полученные гистограммы показывают ненормальное распределение и позволяют сделать следующие выводы:

1) Среднее кол-во посетителей по ресторанам примерно одинаковое без аномальной разницы по пикам.

2) Ресторан 4 выделяется на фоне остальных более стабильным кол-вом посещений. Можно сделать вывод, что этот ресторан пользуется большим спросом.

Для понимания зависимости посещаемости ресторанов от дней недели построим следующие графики (рисунок 3).

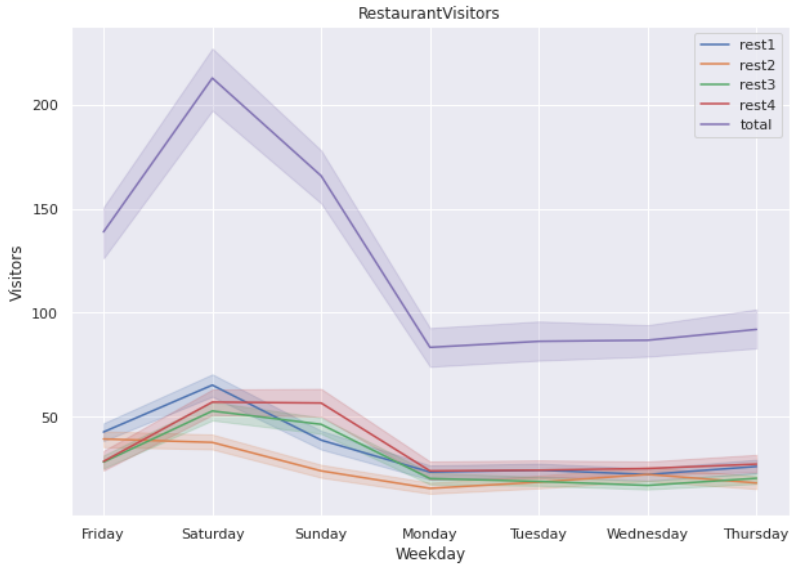


Рисунок 3 – Зависимость посещаемости ресторанов от дня недели

Для выявления корреляции посещаемости ресторанов от дней недели построим следующие виды графиков (рисунок 4).

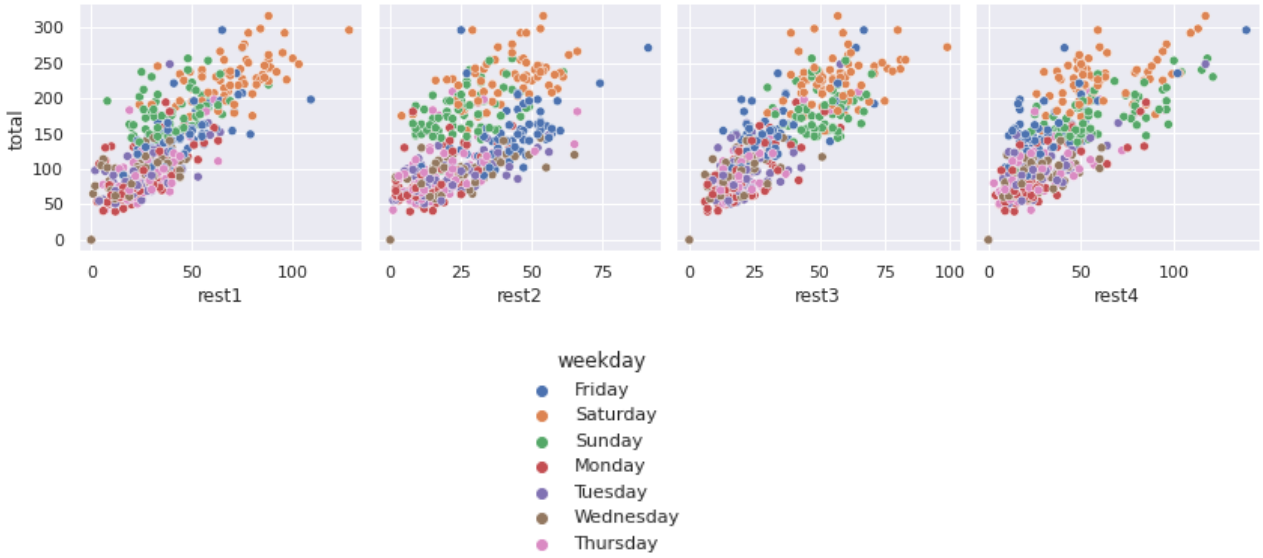


Рисунок 4 – Зависимость посещаемости каждого ресторана от общего кол-во посещаемости по дням недели

По данным графикам прослеживается корреляция посещаемости от дня недели. Видно, что рост посещаемости приходится на пятницу-воскресенье с пиком в субботу. Таким образом задача 1 подтверждена.

Построение моделей, анализ результатов

На данном этапе проведем исследование по второй поставленной задаче - прогнозирование общей посещаемости ресторанов (метрика total). Здесь нам потребуется:

- определиться с кол-вом прогнозируемых дней, возьмем 2 недели для более наглядного результата прогноза;

- разделить наш датасет на тренировочную и тестовую выборки, тренировочную используем для обучения моделей, а тестовую для проверки точности выполненного прогноза;

- определится с моделями прогнозирования, здесь выберем 3 модели - в качестве основной используем SARIMAX, поскольку у нас явно выраженная сезонность, также проверим на нашем датасете работу моделей Prophet и Экспоненциальное сглаживание для выявления возможно более точных результатов предсказаний.

- в завершении обучим наши модели, получим прогнозные данные и сравним с тестовой выборкой.

На предыдущих этапах был выявлен "хвост" с нулевыми значениями - 39 строк в конце датасета для атрибутов/метрик rest1-4 и total. Для корректной выборки данных обучающей и

тестовой групп, и как следствие для более реального прогноза, создадим новый датасет не включающий последние нулевые строки.

Выводы по модели SARIMA (рисунок 5):

- Полученные результаты являются удовлетворительными:

- По результатам видно примерное соответствие графиков прогнозной и тестовой выборок.

- Средняя абсолютная ошибка в процентах = 18%, что также достаточно неплохо.

Модель предсказывает сохранение сезонности и стабильной посещаемости в последующие 2 недели.

Модель может быть использована в работе для дальнейшего прогнозирования.

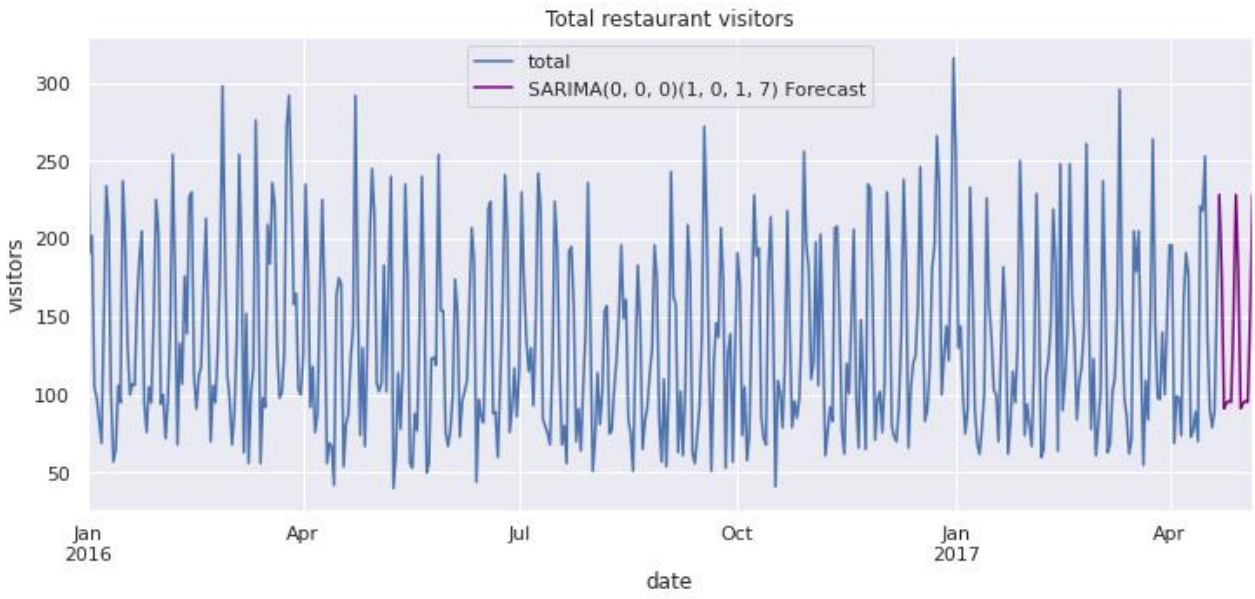


Рисунок 5 – Общий график с исходными и прогнозными значениями на 14 дней

Выводы по модели Prophet (рисунок 6):

Полученные результаты в целом хуже чем по модели SARIMA:

- Средняя абсолютная ошибка в процентах = 64%, против 18% по SARIMA.

- При этом модель также обнаруживает сезонность.

Использование данной модели требует дополнительного изучения.

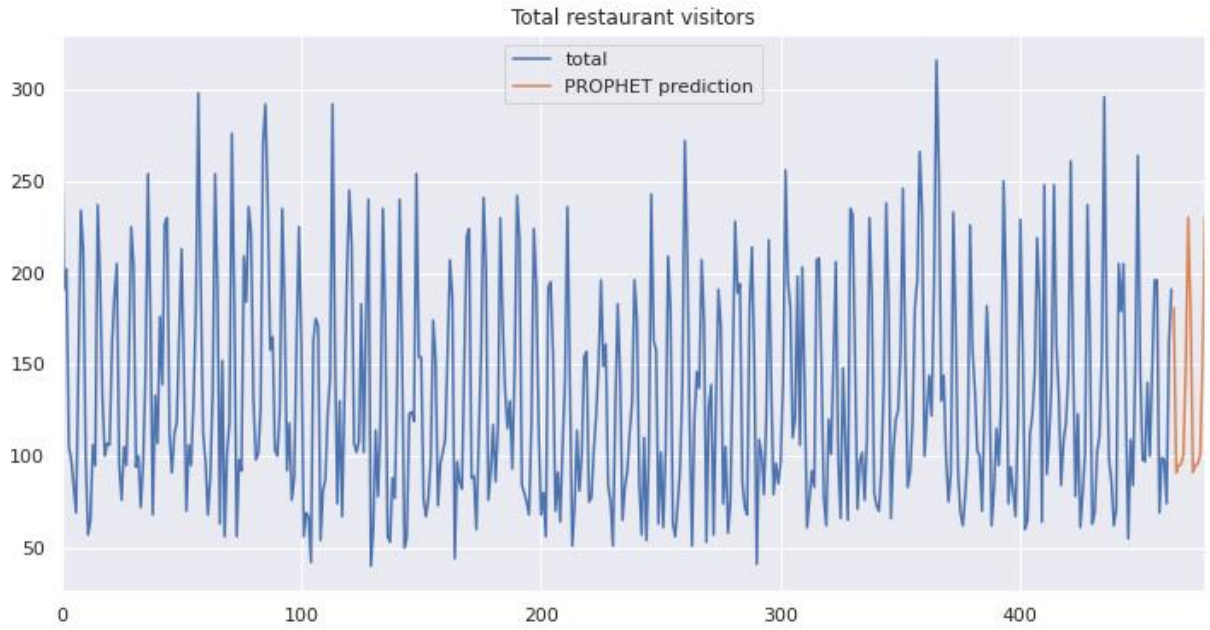


Рисунок 6 – Общий график с исходными и прогнозными значениями на 14 дней

Выводы по модели "Экспоненциальное сглаживание" (Exponential smoothing) (рисунок 7):

Полученные результаты также в целом хуже чем по модели SARIMA:

- Средняя абсолютная ошибка в процентах = 48%, против 18% по SARIMA.
- При этом модель не показывает сезонность.

Использование данной модели для данного датасета нежелательно.

Рисунок 7 – Общий график с исходными и прогнозными значениями на 14 дней

Общие выводы:

Для решения поставленных задач, выявить зависимости посещаемости от дней недели и спрогнозировать посещаемость на несколько дней вперед, были выполнены следующие шаги:

- Проведена обработка и фильтрация исходных данных из полученного датасета;

- Произведен расчет и анализ статистических данных по основным атрибутам;

- Определены основные метрики (посещаемость ресторанов), построены графики временных рядов для данных метрик;

- Проверена корреляция основных метрик;

- Изучены прогнозы на 14 дней по 3 моделям (SARIMA, Prophet и Exponential smoothing).

По результатам анализа можно сделать следующие заключения:

- Подтверждена корреляция между изменением посещаемости и днями недели. Так на предоставленных данных подтвержден известный факт, что в выходные дни посещаемость значительно выше чем в будние - отмечается рост посетителей начиная с пятницы и по воскресенье. Пиковая посещаемость приходится на субботу.

- Среди изученных моделей лучшее качество прогнозирования показала модель SARIMA. Здесь были получены минимальные ошибки прогнозов по сравнению с моделями Prophet и Exponential smoothing. Поэтому для дальнейшей работы можно использовать именно эту модель.

- Результаты прогнозов положительные с сохранением тенденций на сезонность и стабильную посещаемость.