Time series analysis: The change in the business of restaurants

Before and after COVID-19

Patrick Costa[1] Matteo Mondini[2]

[1]University of Milano-Bicocca, E-mail: p.costa7@campus.unimib.it, Matr: 858156
[2]University of Milano-Bicocca, E-mail: m.mondini11@campus.unimib.it, Matr: 902873

Abstract

The proposed project entails a thorough examination of sales trends across six restaurants in the Lombardy and Emilia-Romagna regions, covering the period from September 2018 to April 2023. Our objectives are threefold: first, to discern significant patterns in historical sales data; second, to calculate the financial impact of closures due to the pandemic; and third, to predict future performance for these establishments. The project begins with an exploratory analysis of the collected historical sales figures from the selected restaurants. The purpose of this phase is to detect and interpret various patterns, including seasonal variances and general trends across the establishments. Subsequently, we will utilize a suite of models to rigorously analyze time series data, which will also include an assessment of potential losses stemming from the COVID-19 pandemic. Lastly, these models have been applied to estimate the future sales for the year 2023, for which data is presently unavailable.

Keywords: Sales trend analysis, pandemic impact quantification, time series forecasting.

Contents

1 Introduction

Over the recent years, the COVID-19 pandemic has wrought significant challenges for businesses globally through the enforcement of restrictive measures aimed at curtailing the virus's spread. The pandemic's first documented case was reported in China in early January 2020, but its effects began to be acutely felt in Italy by the latter half of February of the same year. The progressive tightening of restrictions led to initial closures of educational institutions, ultimately escalating to a comprehensive lockdown initiated on March 9. The subsequent two years were marked by a series of intermittent openings and closures, with the degree of restrictions being modulated in response to the shifting intensity of the pandemic.

The restaurant industry was particularly hard-hit, with the onset of closures followed by a phased and often constrained reopening, accompanied by significant shifts in consumer habits. These disruptions have led to substantial economic losses for the industry, and in certain cases, the unfortunate permanent cessation of operations for some establishments.

In response to these widespread disruptions, our project commences with a preparatory phase of data pre-processing. This involves augmenting the initial dataset with additional context to enrich the analysis of historical sales data. We then conduct an exploratory examination of the data, scrutinizing sales trends across various restaurants. As we progress, the project implements a range of models specifically designed for time series analysis. The objective of these models is to trace the sales trajectory throughout the pandemic and to extrapolate these findings for future sales forecasting. Our overarching goal is to delineate the complex relationship between the pandemics impact and the future outlook of the restaurant industry.

2 Research Questions

The research endeavors to answer the following questions:

1. What are the main characteristics of the different restaurants, and can exploratory analysis identify significant patterns within the historical sales data available?

2. Considering the impact of the COVID-19 pandemic, what would have been the expected sales levels for various restaurants if the pandemic had not occurred, and is it possible to quantify financial losses during the pandemic using various time series analysis models?

3. What are the possible sales predictions for the restaurants in the months following April 2023, and which model is the most reliable for estimating future sales levels?

3 Data

The dataset, provided by the University of Milano-Bicocca, encompasses sales data for six restaurants located in Northern Italy, specifically in the regions of Lombardy and Emilia-Romagna, and the cities of Piacenza, Montebello della Battaglia, Voghera, and Stradella. The initial dataset comprises four attributes:

- **Date:** This attribute records the date corresponding to each entry, formatted as *yyyy-mm-dd*.
- **Receipts:** A discrete quantitative attribute indicating the number of receipts issued on each respective day.
- **Total gross:** This attribute details the daily gross earnings.
- **Restaurant ID:** A qualitative attribute representing the unique reference code assigned to each restaurant.

3.1 Data integration

The integration of the initial dataset involved the incorporation of supplementary attributes, as follows:

- **location, region, and province:** These attributes were generated based on the restaurant ID and matched with the provided information regarding the restaurants geographical location.
- **day, month, and year:** Qualitative attributes were derived from the existing date attribute, providing finer temporal granularity.
- **Weekend:** A boolean attribute, assuming a value of True for Fridays, Saturdays, and Sundays, and False otherwise.
- **Holiday:** A boolean attribute, taking on True if the day corresponds to a holiday, encompassing Sundays, nationally recognized holidays, and local patron saint festivities; otherwise, False. Each day of holiday is found for each specific restaurant, based on the location.
- **Color:** A qualitative attribute indicating the color assigned to the region as a measure of COVID-19 containment. The color is determined by the region to which the restaurant belongs and the specific day under consideration. "Red" and "Orange" shows that there was a lot of restrictions and the restaurants were closed, "Yellow" and "White" had way fewer restrictions and the restaurants were open to the public.
- **Lockdown:** Lockdown period (0 or 1 - no/yes).
- **Season:** The season in which the day falls.
- **TAVG (Air temperature):** TAVG represents the average air temperature recorded over a day, measured in degrees Celsius .
- **TMIN (Minimum temperature):** TMIN is the lowest air temperature recorded during a day.
- **TMAX (Maximum temperature):** TMAX is the highest air temperature recorded during a day.
- **PRCP (Total precipitation):** PRCP indicates the total amount of precipitation (rain, snow, sleet, or hail) that occurred during a day, measured in millimeters .

- **WDIR (wind (from) direction):** WDIR tells us the direction from which the wind is blowing during the day, expressed in degrees.
- **WSPD (average wind speed):** WSPD is the average wind speed recorded during the day, commonly measured in units like kilometers per hour (km/h).

4 Explorative analysis

4.1 Analysis of gross sales amount in general

In the pursuit of a nuanced understanding of the dynamics within the restaurant sector, we present an elaborate comparative analysis of the gross sales trends, encapsulated by two distinct yet interconnected metrics: the daily mean and sum gross amounts. These metrics offer a dual perspective on the financial throughput of the six restaurants in question, before the deduction of any operational expenditures.

Figure 1 delves into the average revenue per transaction, charted daily from September 2018 through May 2023. This visualization, through its methodical delineation of the mean gross amount, encoded in a vivid spectrum of blue, articulates the rhythmic financial heartbeat of day-to-day operations. The graph's temporal journey is punctuated by the strategic application of moving averages, with the 7-day and 30-day moving averages rendered in variegated shades of green and orange acting as analytical sentinels that distill the intrinsic sales volatility into a coherent narrative, revealing the steady undercurrents of consumer spending patterns.

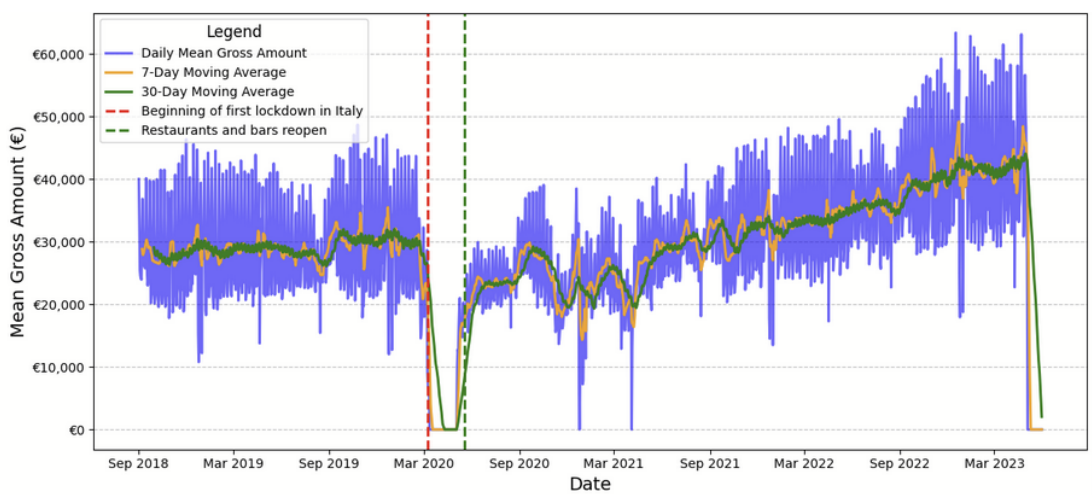

Figure 1: Daily mean gross amount with moving averages between September 2018 and May 2023, mean across all the restaurants.

In juxtaposition, Figure 2 shifts the focus from individual transactions to the collective fiscal magnitude, presenting the sum gross amount over the same temporal expanse. This composite summation transcends individual performance, showcasing the aggregate revenue prowess of the collective entities. Maintaining the aesthetic continuity of the preceding figure, this chart communicates a parallel financial saga, chronicling the cumulative sales volumes and their broader economic ramifications.

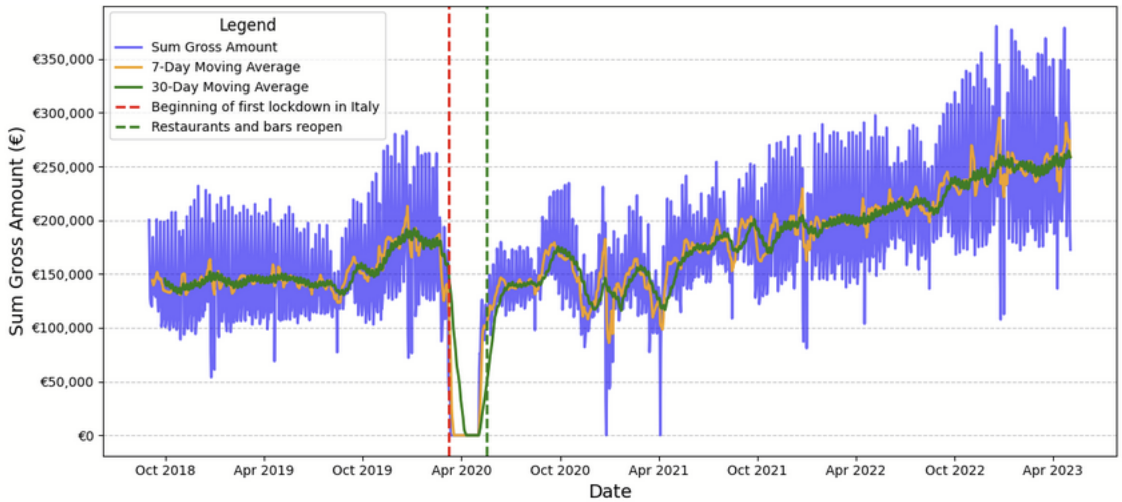

Figure 2: Daily gross amount with moving averages between September 2018 and May 2023, sum across all the restaurants.

Both figures are strategically annotated with temporal signposts that demarc-ate seminal events in recent history. The 9th of March 2020 is etched into the charts with a red dashed vertical line, signifying the implementation of Italy's initial lockdown, a response to the COVID-19 pandemic. The pronounced downward trajectory in sales following this demarcation serves as a visual testament to the pandemic's immediate and profound financial impact on the restaurant industry.

Emerging from the shadow of the lockdown, the green dashed line heralds the reopening of restaurants and bars, marking a pivotal turning point. This milestone captures the industry's gradual resurgence, charting a path of economic recovery that underscores the sector's fortitude and the rekindled consumer engagement in the aftermath of the restrictions.

The intricate interplay between the daily mean and sum gross sales amounts, as laid bare by these figures, provides a comprehensive financial overview. It underscores the sector's vulnerability to external shocks, yet simultaneously highlights its capacity for rapid recalibration and recovery. These insights, derived from a meticulous synthesis of granular sales data and moving averages, are indispensable for stakeholders seeking to formulate strategic responses to fluctuating market conditions.

After a general analysis, we can delve into a more specific view to see the details of the single restaurants. In terms of general trend variability, the financial performance of R000 3 is notably volatile.

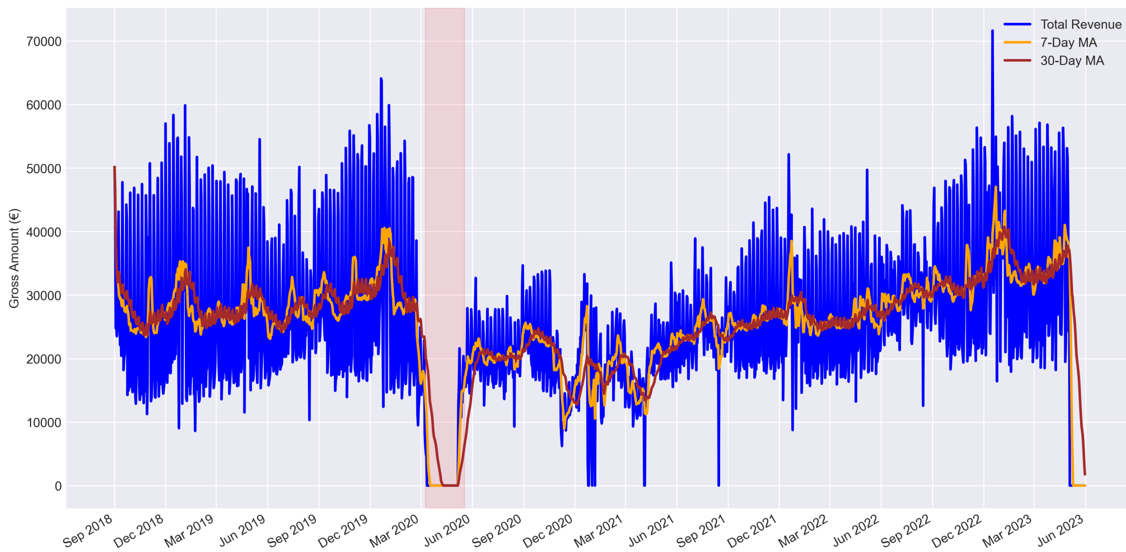

Figure 3: Daily gross amount with moving averages for restaurant R000(sep 2018-may 2023).

This restaurant's earnings chart paints a picture of frequent oscillations with stark highs and lows, making its revenue streams unpredictable. On the other hand, restaurants R001 in fig.4 and restaurant R004 in fig.5 exhibit a more tempered pattern.

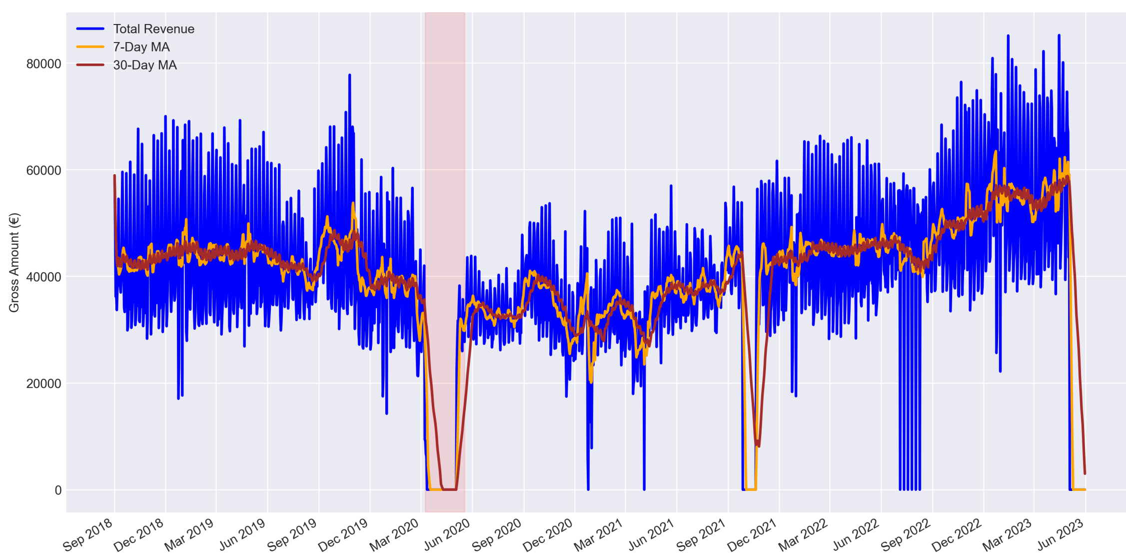

Figure 4: Daily gross amount with moving averages for restaurant R001(sep 2018-may 2023).

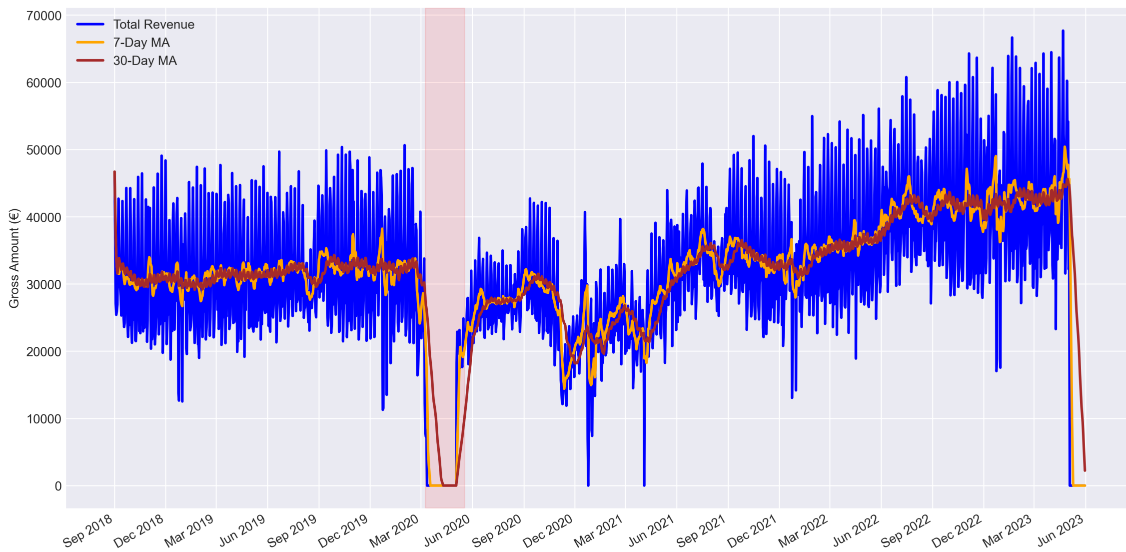

Figure 5: Daily gross amount with moving averages for restaurant R004(sep 2018-may 2023).

While they too navigate the ebb and flow inherent in the industry, their fluctuations are less dramatic than those of R000, suggesting a certain level of consistency in their operations or customer demand. Conversely, R002, R003, and R005 appear to be the paragons of stability among the group. Their revenues, while still subject to change, do so with a degree of predictability and lack the wild undulations seen in R000.

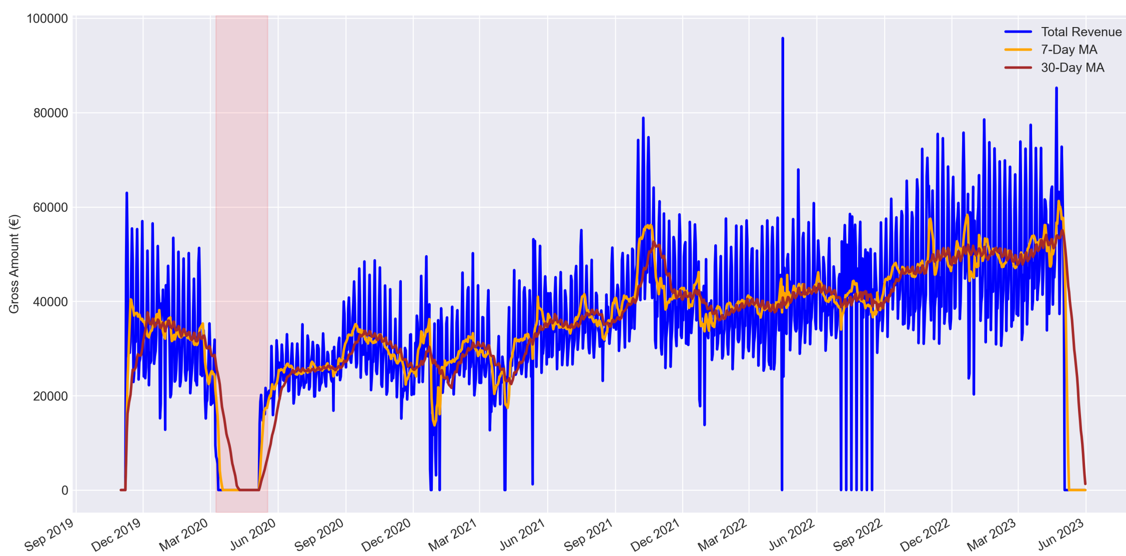

Figure 6: Daily gross amount with moving averages for restaurant R002(sep 2018-may 2023).

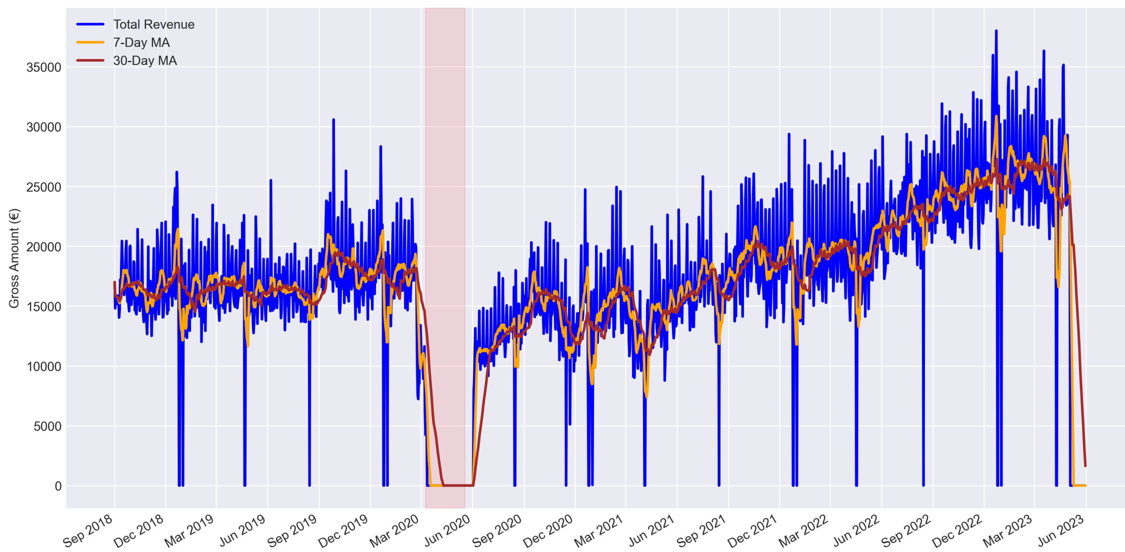

Figure 7: Daily gross amount with moving averages for restaurant R003(sep 2018-may 2023).

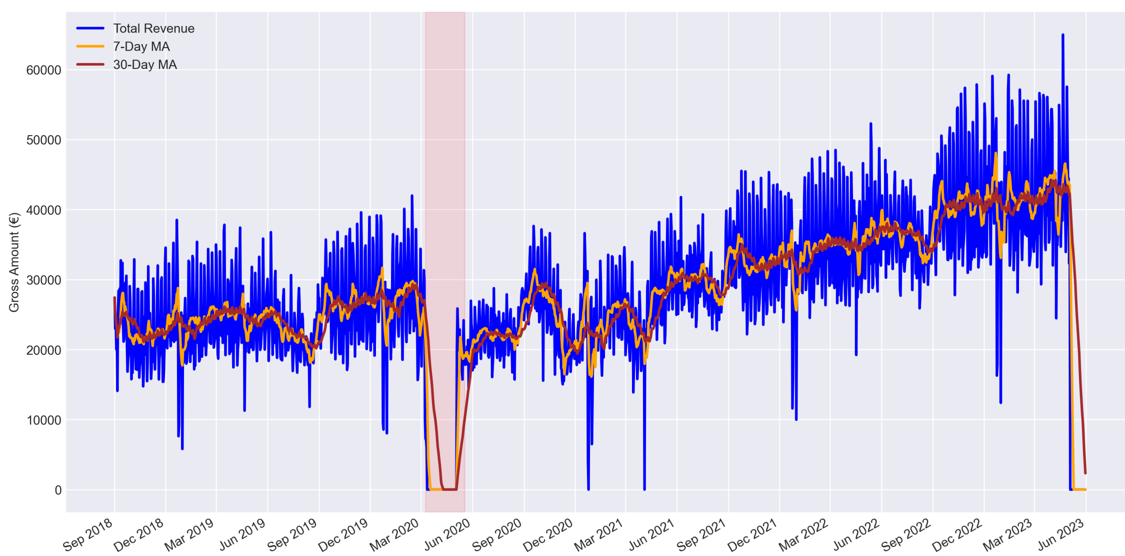

Figure 8: Daily gross amount with moving averages for restaurant R005(sep 2018-may 2023).

The global lockdown brought about by the COVID-19 pandemic undeniably left its mark on the hospitality sector. R001's resilience during this period is commendable. Despite facing significant economic challenges during the lockdown, it exhibited an impressive ability to bounce back , nearly mirroring its pre-lockdown earnings in a relatively short span. In contrast, as we can see the restaurants R002 in fig.6 and R003 in fig.7 underwent more measured recoveries. Their post-lockdown trajectories were characterized by a steady, albeit slower, climb back to their previous revenue levels. R004 in fig.5 and R005 in fig.8 charted a path of consistent growth in the aftermath of the lockdown, while R000's recovery seemed somewhat muted, hinting at potential challenges in adapting to the post-lockdown landscape. When reflecting on the earnings potential of these establishments, R001 stands out as the top earner. It regularly achieves and occasionally even breaches the 80,000 mark, setting a high benchmark for its peers. R002 also demonstrates robust earning capabilities, with its revenue peaks touching the 80,000 threshold at times. R000 and R005, in comparison, seem to operate on a more modest scale. Their earnings typically hover around 60,000. However, a silver lining emerges in their post-pandemic recovery, suggesting an admirable ability to adapt and rally after a global crisis.

The chart in fig. 9 provides a detailed visualization of the average weekly sales for all the six restaurants. The temporal evolution of sales is captured through individual lines, each representing one restaurant and colored distinctly to differentiate their respective sales patterns.

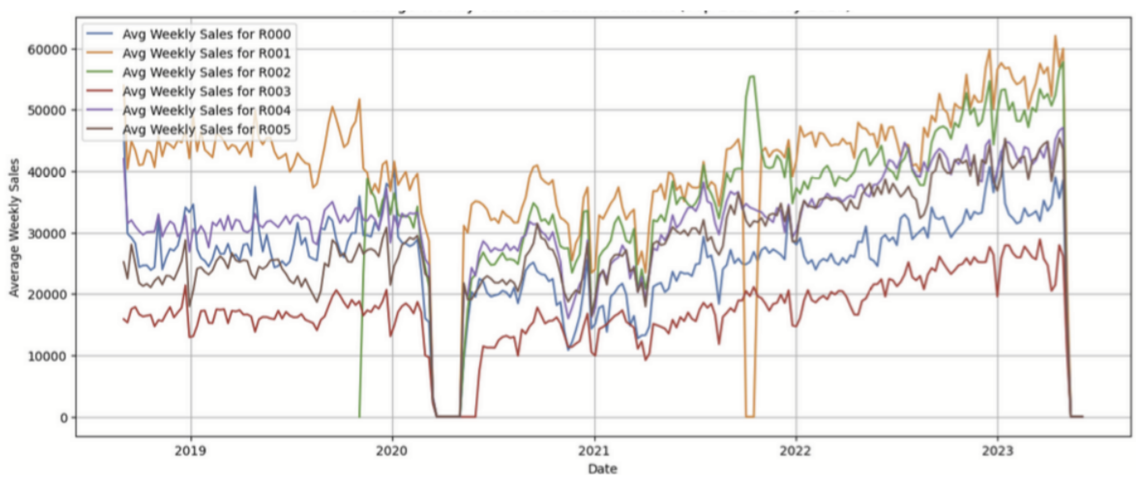

Figure 9: Average weekly sales for each restaurant(sep 2018-may 2023)

The other chart in fig. 10 shifts the perspective to a comparative annual analysis, plotting the average weekly sales across the years for each restaurant. This comparative approach highlights the annual performance and potential growth or decline in customer patronage. Restaurants such as R001 and R004 display a consistent upward trend, indicating a solid year-over-year growth and possibly effective strategic initiatives. In particular, the resilience of R001 post-2020 is noteworthy, showcasing a rapid recovery and a return to pre-pandemic sales levels, suggesting robust business continuity plans.

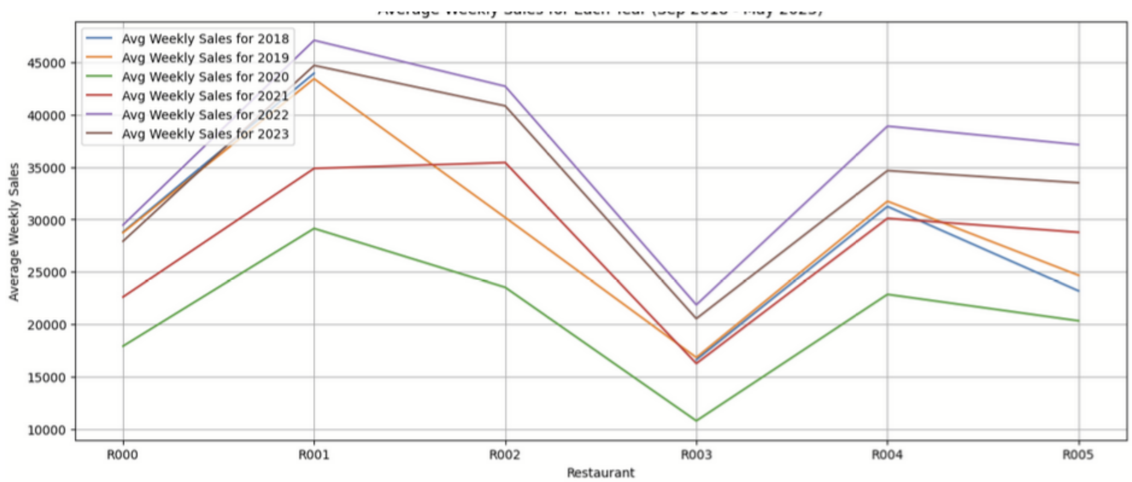

Figure 10: Average weekly sales for each restaurant(sep 2018-may 2023)

On the other hand, R002 and R003 exhibit a more gradual recovery post-lockdown, with a steady climb back to their pre-pandemic performance. This may reflect a cautious approach to resuming operations or a gradual return of their customer base. R004 and R005 show a consistent pattern of growth post-lockdown, possibly benefiting from shifting market dynamics or successful adaptation to the new normal.

In contrast, R000's recovery appears more subdued, indicating potential challenges in adapting to post-pandemic market conditions or other operational hurdles.

4.2 Analysis of receipts

The temporal dynamics of customer spending in the restaurant sector are comprehensively depicted in fig.11. This figure presents the progression of daily average receipt prices across a consortium of restaurants over the analyzed period.

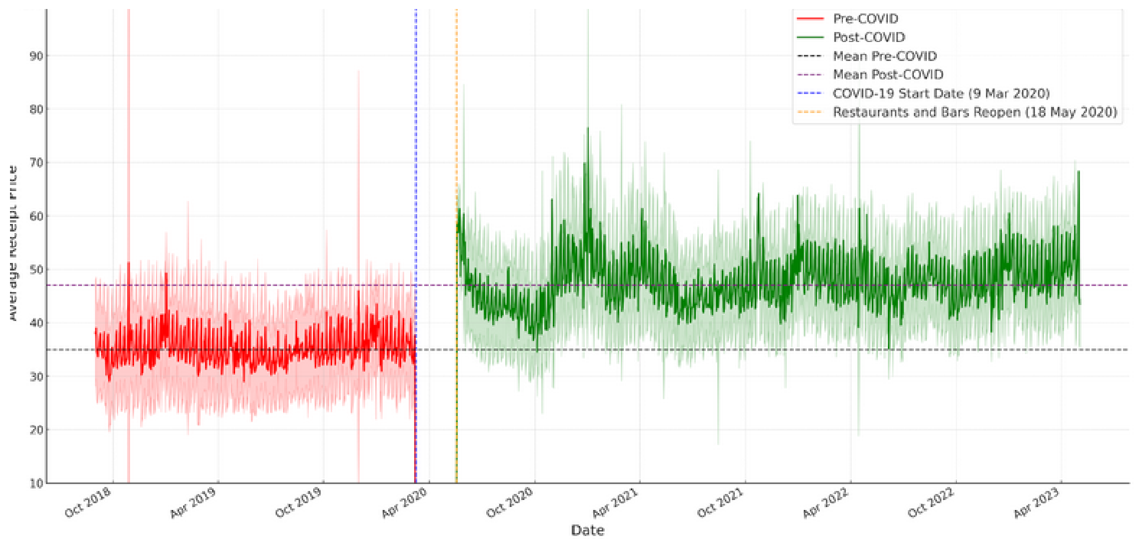

Figure 11: Daily average receipt price of all the restaurants during the period sep 2018 - may 2023

A discernible bifurcation in the data corresponds to the pre- and post-COVID-19 periods. In the pre-pandemic landscape, the average receipt price oscillates with relative uniformity around a central mean, suggesting a period of economic stability within the sector. The regularity of these fluctuations is indicative of typical consumer spending behavior, unimpacted by the externalities of the impending health crisis.

The advent of the COVID-19 pandemic, demarcated by a vertical dashed line in the figure, signifies a pivotal trend in consumer expenditure patterns. This disruption is visually represented by a clear deviation from the pre-established mean, leading in a period of pronounced economic uncertainty. Subsequent to this delineation, a second vertical line marks the advent of a gradual resumption of normalcy, as restaurants and bars commence reopening.

Post-reopening, a conspicuous elevation in the mean receipt price is observable, potentially reflective of inflationary pressures, alterations in menu prices, or a shift in consumer behavior towards higher spending per visit, possibly due to a pent-up demand following the easing of restrictions. The persistence of volatility post-reopening underscores the sector's ongoing recalibration to the 'new normal' of operational constraints and consumer caution

The analysis affords a nuanced understanding of the financial vicissitudes experienced by the restaurant industry throughout a challenging economic period marked by a global health emergency. It underscores the imperative for robust adaptive strategies to navigate through periods of economic tumult and the critical nature of flexibility within operational models to maintain fiscal viability.

The graph shew in the fig.12 reveals the distinctive trajectories of receipts, underscoring varied customer engagement levels, which may be attributed to the unique attributes of each restaurant, including location, cuisine, and management strategies. Notably, the precipitous declines and subsequent recoveries apparent in the graph correspond with the globally disruptive COVID-19 pandemic, elucidating the profound impact it has had on the hospitality sector.

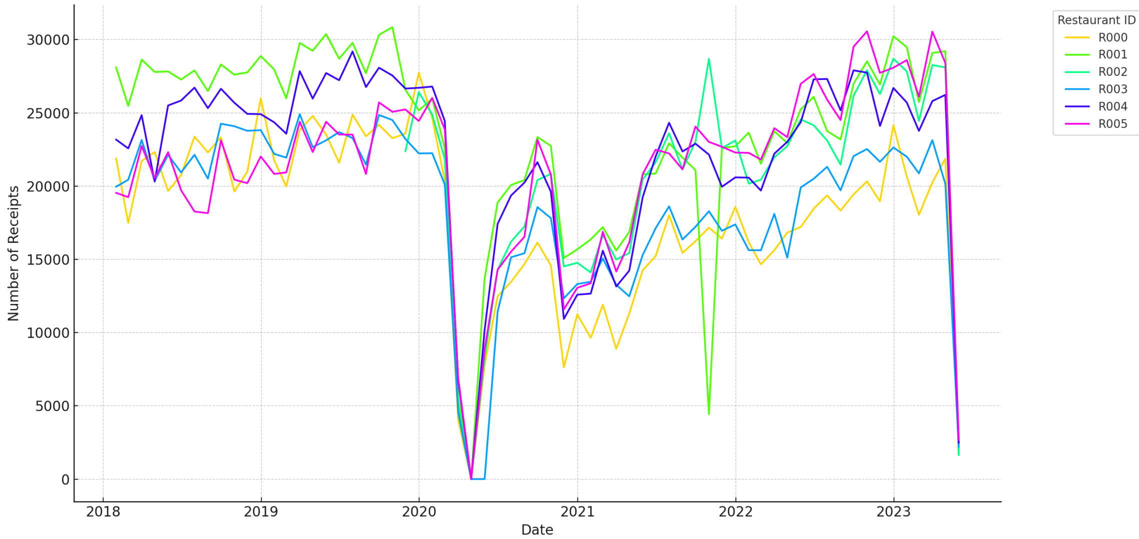

Figure 12: Montly receipt number for each restaurants during the period sep 2018 - may 2023

The analytical focus on individual entities, rather than an aggregated overview, allows for a granular assessment of business health and consumer behaviour. These insights are very important for creating tailored strategies that meet the unique challenges and opportunities of each restaurant.

4.3 Analysis for different factors

In this part of the analysis, we looked at how sales were influenced by certain attributes that were added during the data integration phase. To achieve this goal, we created box plots that showed the sales data divided based on different categorical variables. We specifically looked at the time before and after the start of the pandemic to see if there were any noticeable differences in behavior. These box plots helped us identify any inconsistencies or changes in sales patterns. By comparing the sales data before and after the pandemic, we were able to draw some interesting conclusions about how different attributes affected sales during these periods.

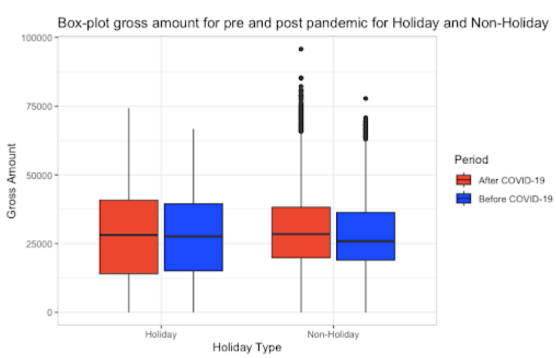

Figure 13: Box-plot Holidays

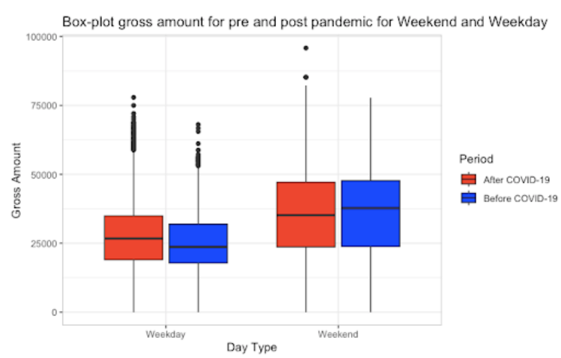

Figure 14: Box-plot Weekend

In Fig.13, the sales data for six restaurants from september 2018 to end may 2023 shows a noticeable distinction between Holiday and non-Holiday periods. Generally, during Holidays, there's a widespread increase in sales variability across restaurants. While some establishments experience higher median sales, the overall pattern indicates a greater fluctuation in sales during festive seasons.

Contrastingly, non-Holiday periods are marked by more prominent sales outliers, suggesting sporadic and substantial sales spikes outside traditional holiday times. This could be attributed to various factors, such as localized events, promotions, or unique marketing efforts by the restaurants aimed at boosting sales on regular days.

Moving to Fig.14, the analysis of weekend versus non-weekend sales provides another interesting insight. Generally, there is a consistent trend of heightened sales during weekends across the majority of the restaurants. This may be attributed to increased leisure time and social activities, prompting more people to dine out during the weekends. However, it's essential to note that the impact of weekends on sales varies across the six restaurants, reflecting the influence of factors such as location, target audience, and menu offerings.

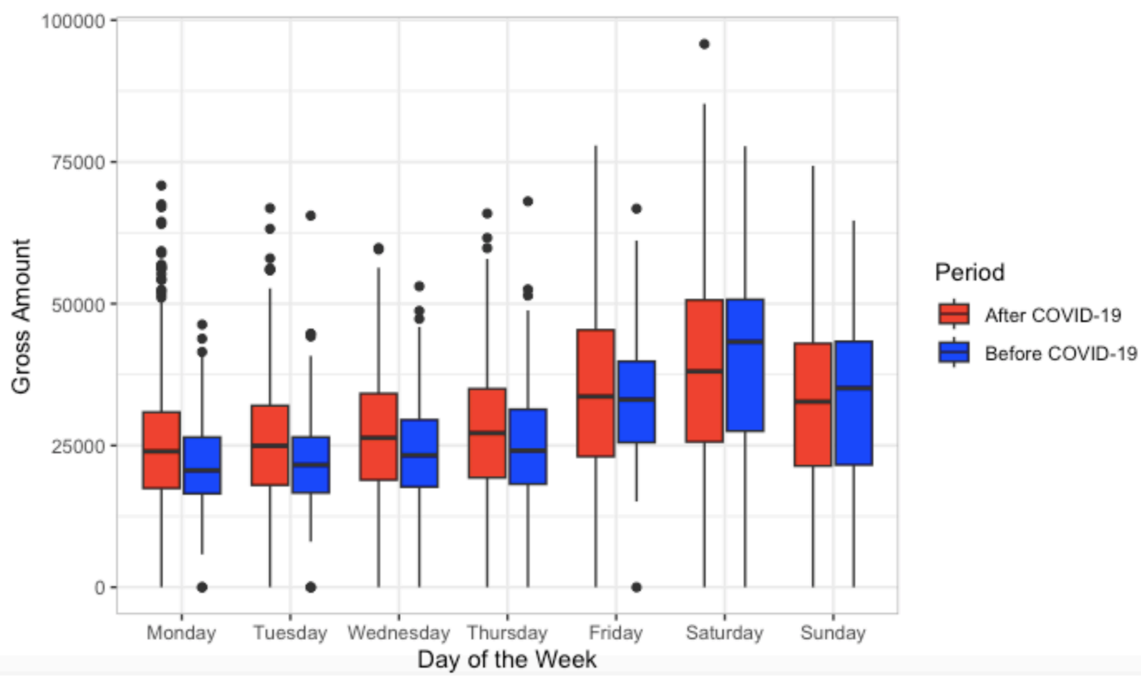

Figure 15: Box-plot weekdays

In Fig.15, the sales data for all the six restaurants spanning 2018-2023 is presented for each day of the week. Although there is variability among the restaurants, certain days consistently stand out with higher sales for particular establishments. Moreover, the presence of pronounced sales outliers on certain days suggests the possibility of occasional promotions or special events specific to those days.

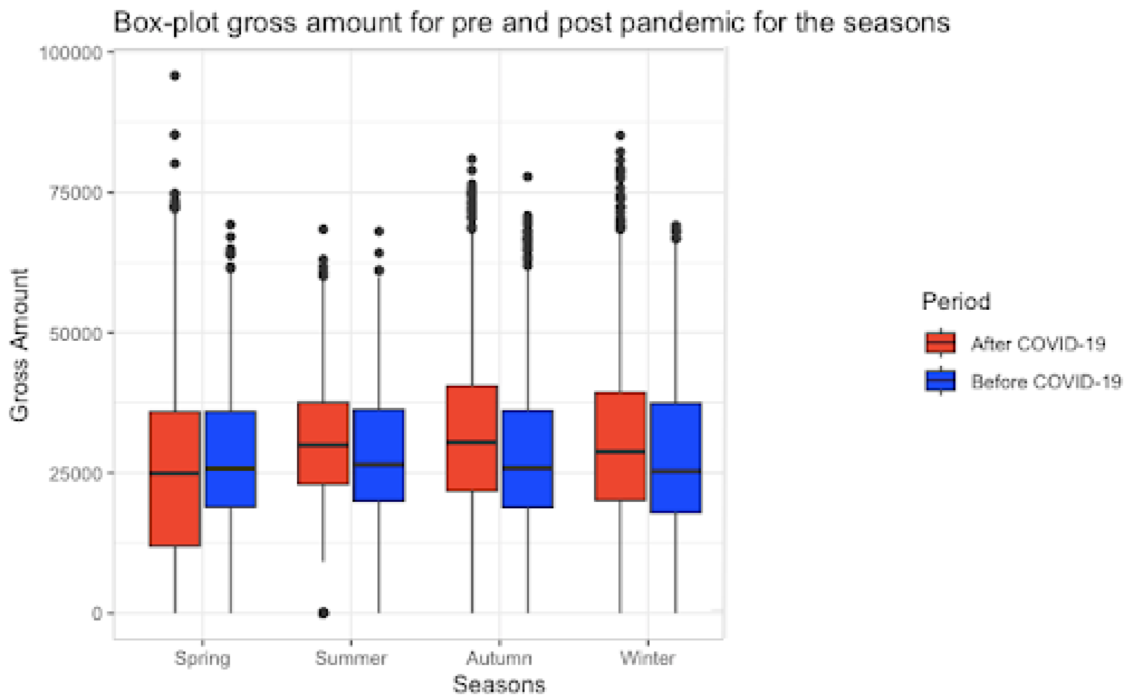

Figure 16: Box-plot seasons

In Fig.16, the sales analysis for the same six restaurants over 2018-2023 is segmented by seasons. The data illustrates distinct variations in sales patterns for each establishment across different seasons. Notably, some restaurants experience heightened sales during specific seasons, hinting at seasonal preferences or targeted promotional activities. The presence of outliers across seasons indicates occasional sales spikes, likely attributed to special events that coincide with specific seasons.

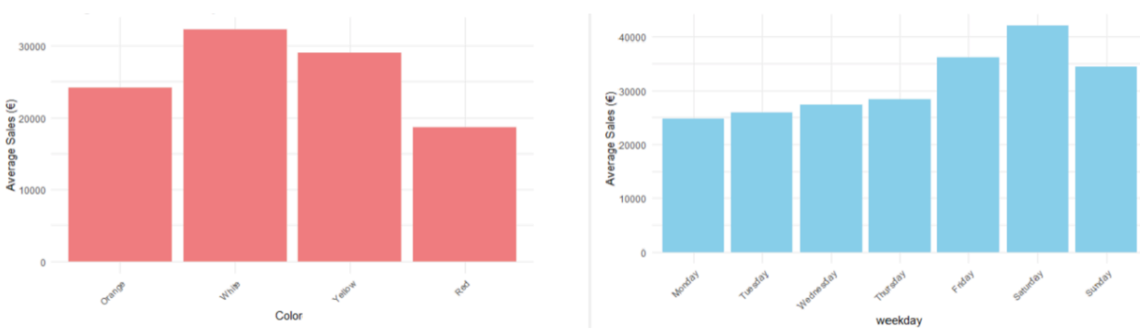

Figure 17: Average Sales distribution by COVID color zones on the left, and weekdays in comparison with average sales on the right

In Fig.17 o the left, the bar chart depicts the average sales distribution categorized by 'Color' for the 6 restaurants over the period of 2018-2023. Notable observations include variations in sales performance across different 'Color' categories, implying that external factors, likely associated with the 'Color' classification (such as restrictions or guidelines), significantly impact sales. Particularly, the highest average sales are recorded during 'White' periods, suggesting a potentially more favorable dining environment or fewer restrictions. Conversely, sales dip during 'Red' and 'Orange' periods, possibly indicating more stringent restrictions or other factors negatively affecting restaurant operations during these times.

Moving to the bar-chart on the right, it provides insight into the average sales distribution across various days of the week for restaurants. Notably, there is a conspicuous spike in sales during the weekend, especially on Saturdays. The mid-week period, from Tuesday to Friday, shows consistent sales, indicating a stable customer base during these weekdays. Mondays, however, exhibit a slight drop in sales, a common trend in the restaurant industry where Mondays tend to be slower.
This pattern reflects the influence of customer preferences and dining habits throughout the week.

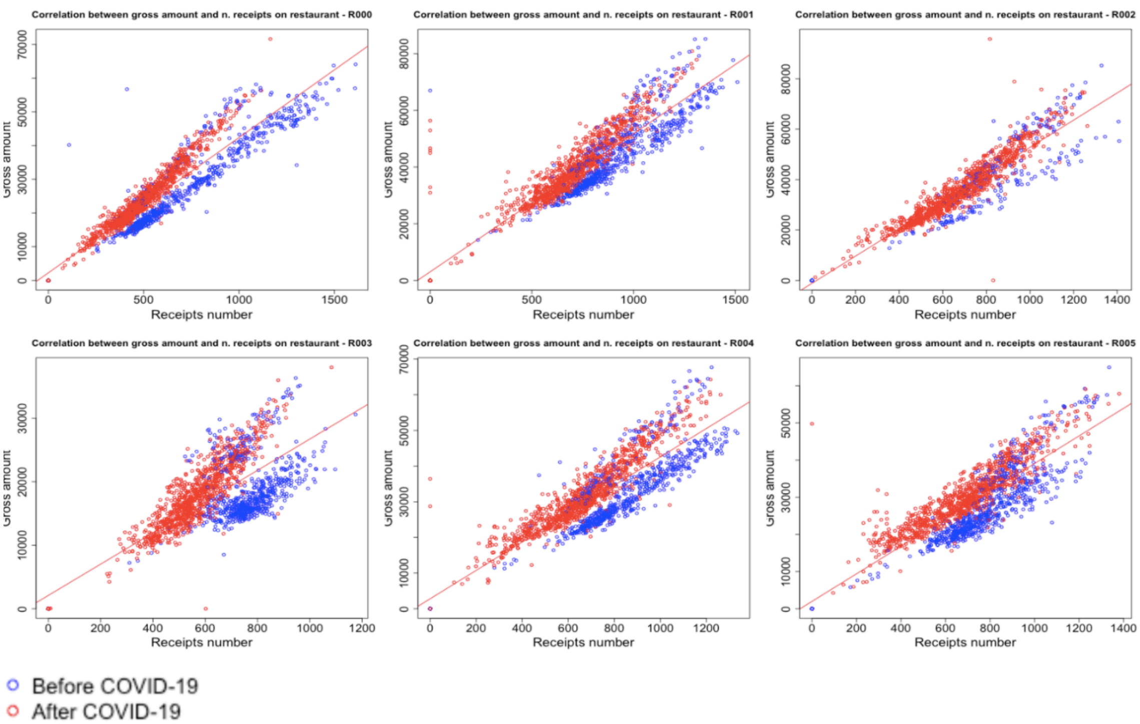

Figure 18: Scatter-plot gross amount in relation with number of receipt before and after COVID

In Fig.18, a clear positive correlation is observed between the number of receipts and total sales for both periods, before and after Covid-19, across the six restaurants. Notably, for R001, R004, R000, and R005, the sales versus receipt dots are slightly higher in the after-Covid period (depicted by red dots), indicating an overall increase in costs. In contrast, for R002, the dots remain relatively consistent, while for R003, the difference is more pronounced, with higher costs observed after Covid-19.

4.4 Analysis of the impact of the climate on profits

Within this section, a comprehensive analysis is undertaken to assess the influence of meteorological conditions on the revenue streams of each individual restaurant. The initial phase of the analysis seeks to discern potential correlations through various data visualizations. Subsequently, an examination of the impact exerted by weather-related variables on sales is conducted, integrating analyses across diverse scenarios. Concluding this multifaceted investigation, a different linear regression model for each restaurant is employed to quantify the effect that variations in weather conditions have on sales.

To preliminarily discern potential correlations between sales and precipitation for restaurants, a scatter plot was constructed (see Fig. 19). This visualization did not reveal any pronounced relationship between the two variables; the distribution of points across the plots appeared stochastic and lacked discernible patterns.

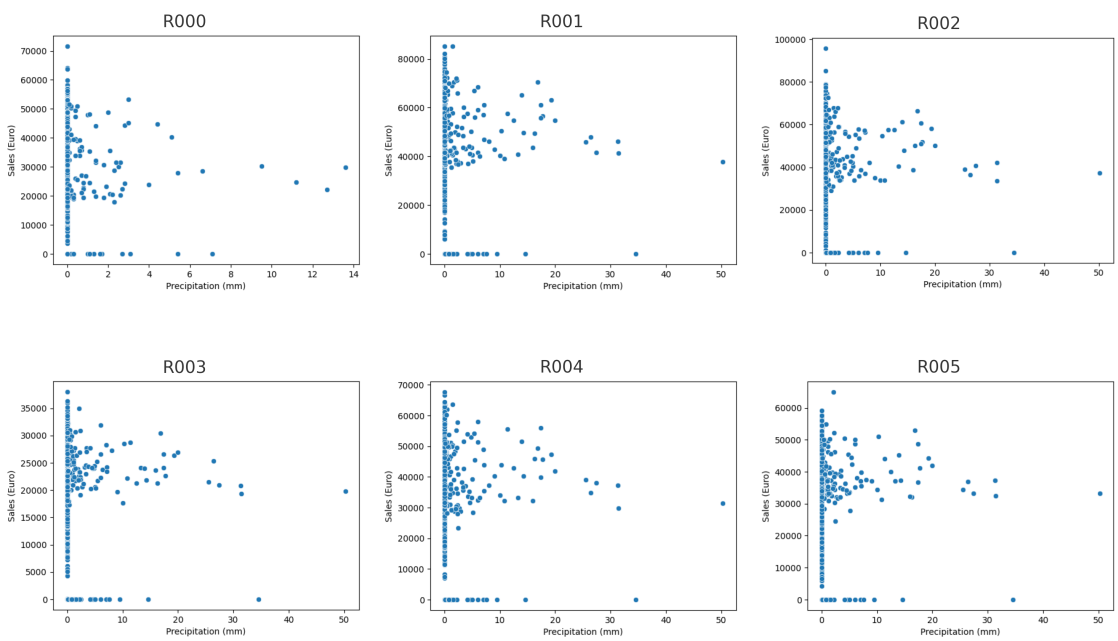

Figure 19: Scatter-plot precipitation and sales for the restaurants

Furthermore, we embarked on an extensive correlation analysis across six different restaurants, aiming to identify key relationships among various variables. This process entailed the exclusion of redundant or interdependent variables, such as maximum and average temperatures, to hone in on more meaningful connections. Interestingly, the findings across these restaurants showed remarkable consistency, suggesting universal trends in the factors influencing restaurant performance.

For illustrative purposes, in in Figure 20. there is the corrplot for the restaurant r004. This example serves as example of the broader trends observed across all six restaurants.

A significant and robust correlation was observed between the number of receipts and total revenue. This relationship underscores a fundamental business principle: an increase in customer count typically leads to higher sales. The consistency of this trend across all restaurants highlights the universal importance of customer volume in the hospitality sector.

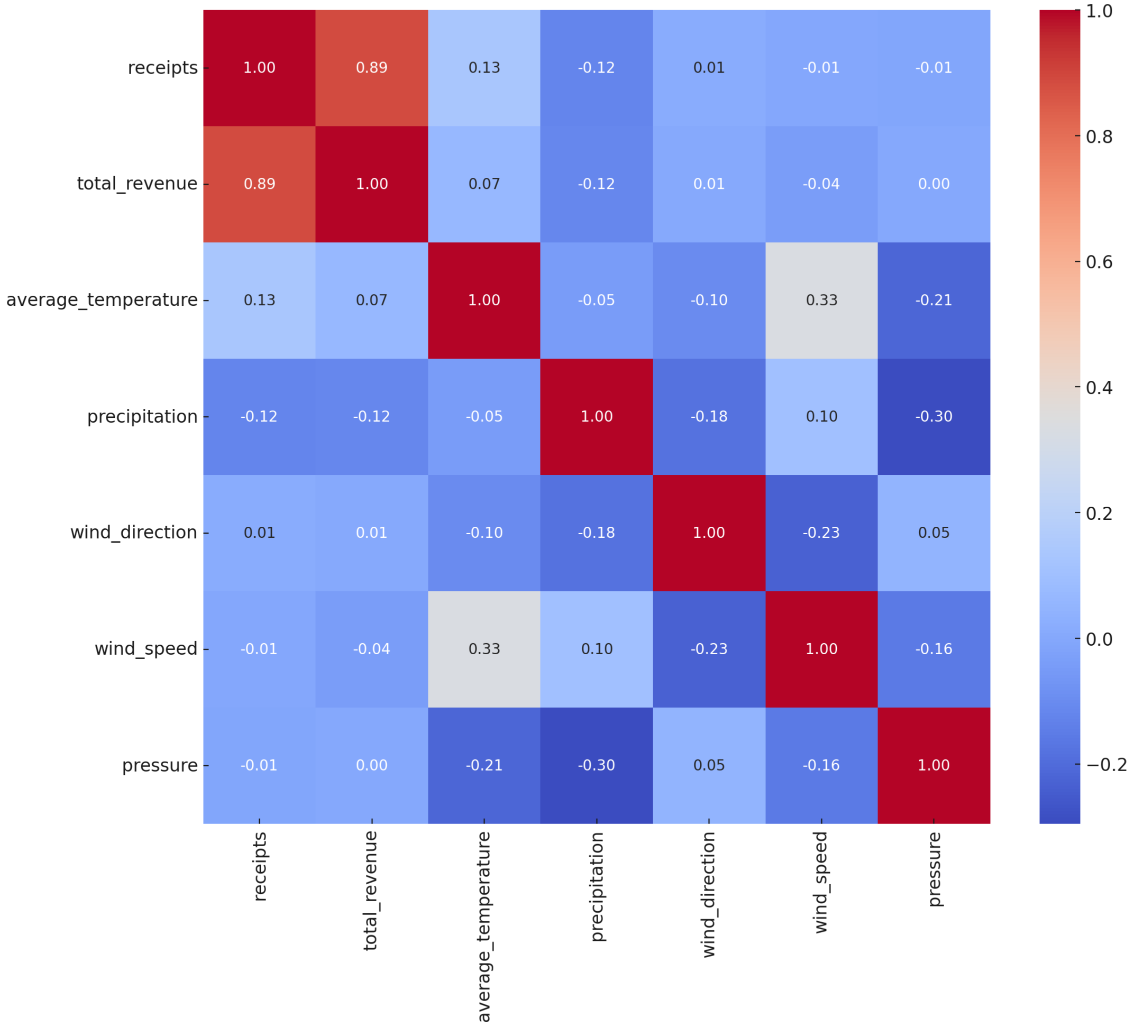

Figure 20: Example of correlation plot of the features of R004

While average temperature showed a correlation with total revenue, its impact was not as pronounced as that of customer numbers. This suggests that while weather conditions like average temperature may play a role in influencing customer decisions, their effect is more nuanced and less direct.

Precipitation, on the other hand, had a discernible but not overly strong correlation with revenue. This indicates that, although adverse weather conditions like rain might affect customer turnout, they are not the primary drivers of a restaurant's success.

Other meteorological factors, such as wind speed and atmospheric pressure, exhibited a relatively weak correlation with total revenue. This finding suggests that these elements, while potentially impacting customer behavior to a certain extent, are not key determinants of a restaurant's financial performa

A further investigation into the impact of precipitation on sales, an analysis categorizing precipitation levels was undertaken. The initial strategy of utilizing percentile-based categories was revised in favor of a more standardized delineation due to the modest dataset, which could not yield statistically robust results through percentile segmentation. The adjusted classification aligns with the conventional guidelines that define 'none' for no rainfall, 'very low' up to 1 mm, 'low' to 2.0 mm, 'moderate' up to 6.0 mm, and 'high' for rainfall exceeding this range. This methodological pivot is substantiated by [9], ensuring that our analytical framework resonates with the broader meteorological criteria.

max width=1center

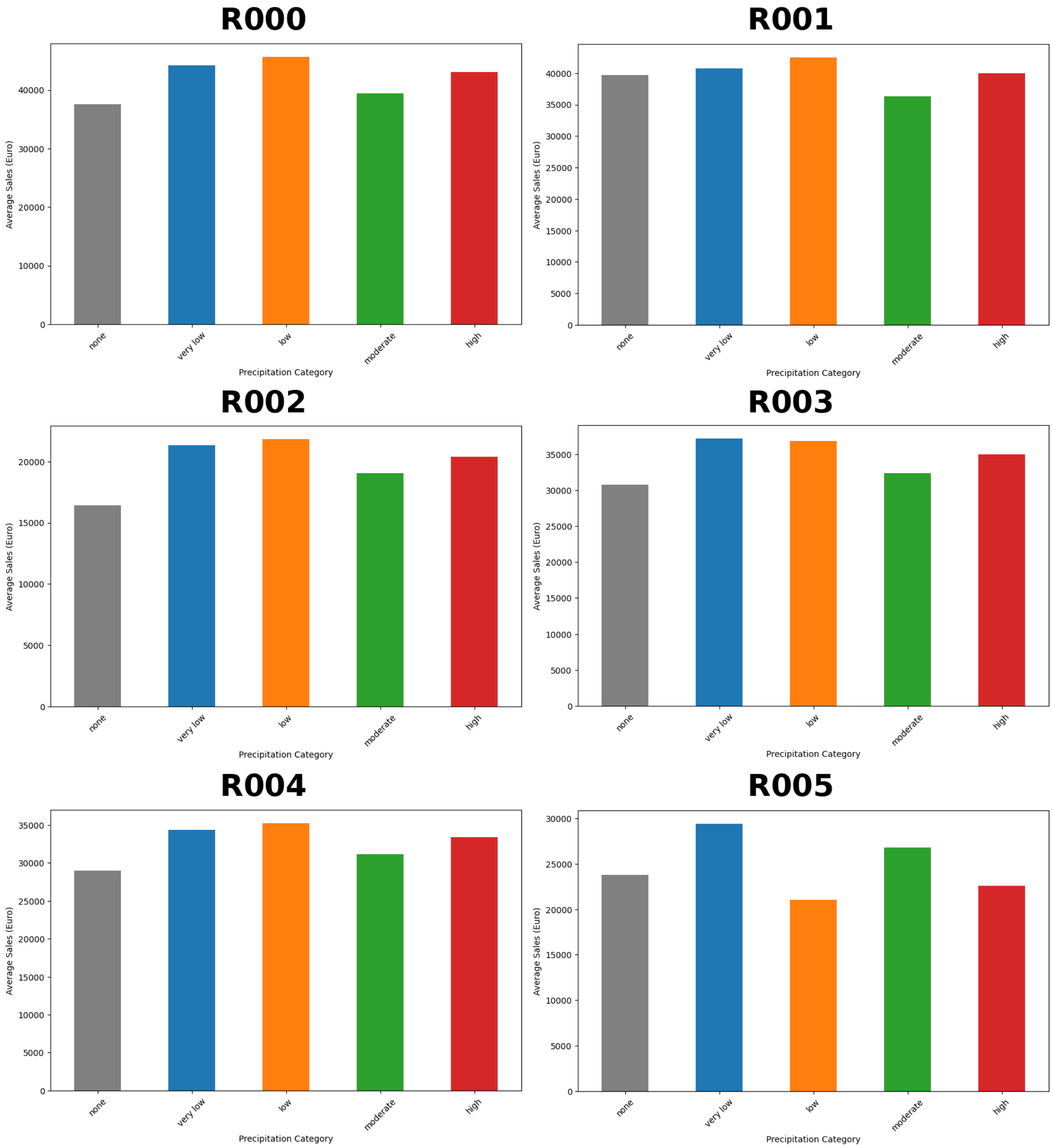

Figure 21: Histograms of average sales based on precipitations for all the restaurants

Subsequent analysis, which purposefully excluded the initial two years marred by a high occurrence of null values, has uncovered a nuanced pattern in consumer behavior. The sales data, illustrated in Fig. 21, does not universally support the hypothesis that average sales are higher on days without precipitation. While this trend holds true for some restaurants, where non-rainy days correlate with increased sales, it is not a consistent pattern across all establishments observed.

In some cases, sales on days with moderate to high rainfall do not significantly lag behind those on dry days, and there are instances where sales during days with 'very low' or 'low' rainfall remain on par with those on non-rainy days. This indicates that the impact of rainfall on sales is more complex than initially suggested and may not necessarily deter customers from dining out.

The variability in sales response to precipitation underscores the influence of other mitigating factors, such as restaurant location, clientele preferences, and the establishment's preparedness for inclement weather. This intricate interplay suggests that while adverse weather can affect consumer behavior, its impact is not homogeneous and is instead modulated by a spectrum of additional variables.

Another comparative study was performed to contrast average sales during extreme weather events against the general average. Extreme weather events were characterized by temperature falling below the 1st percentile or above the 95th percentile, and precipitation exceeded the 95th percentile. These criteria set the stage for evaluating business performance under significantly abnormal weather circumstances.

The results, as depicted in Fig. 22, indicate a marginal discrepancy in sales between extreme weather days and typical days.

Average sales figures during regular weather conditions marginally surpassed those during extreme weather events. This modest dip in sales during extreme weather conditions suggests that while weather factors do play a role, their impact is not as pronounced as might be anticipated. The relative steadiness of sales could be indicative of successful adaptive strategies implemented by the restaurants or possibly a customer demographic that is not significantly deterred by weather extremes.

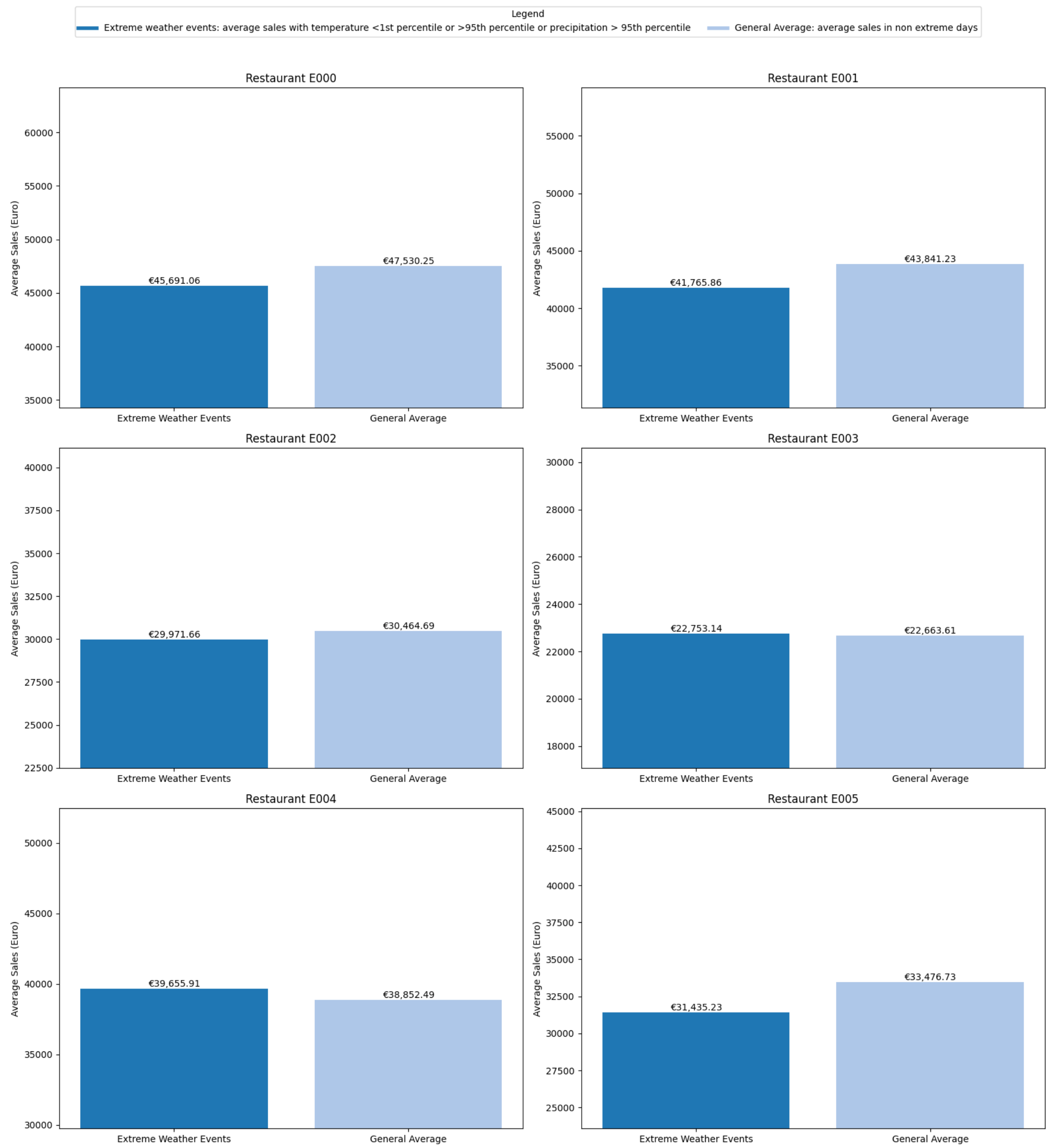

Figure 22: Histogram of average sales of restaurants based on exceptional weather events or ordinary days

It is noteworthy, however, that restaurants R003 and R004 did not follow this general trend, potentially alluding to additional variables at play that may be influencing their sales outcomes. The divergence observed in these establishments warrants a deeper exploration to understand the confluence of factors that may be affecting customer patronage during extreme weather events.

A series of linear regression models were constructed for each time series to quantify the influence of weather variables on restaurant sales. These models incorporated weather conditions and the lagged sales figures as predictors. The analysis of the regression coefficients revealed that temperature (`tavg`) exerted a negligible influence on sales, typically associated with a p-value around 0.5. This indicates that within the data's temperature range, fluctuations failed to produce significant effects on sales.

Precipitation (`prcp`) also had a minimal impact on sales, underscored by p-values exceeding 0.9, suggesting that sales are largely unaffected by variations in precipitation. Conversely, wind speed (`wspd`) displayed a slight negative correlation with sales. Although this relationship was of marginal significance, it hints at the possibility that higher wind speeds might slightly discourage patrons.

During the regression analysis, several variables were excluded due to their trivial contribution to the model and high collinearity with other predictors, such as maximum temperature and atmospheric pressure. The regression models used only the most relevant variables. The R-squared values, peaking at 0.55, indicated that at best, 55% of the variability in sales was accounted for by the models. This reflects only a moderate correlation between the chosen predictors and sales figures.

The lagged sales emerged as a robust predictor, consistently demonstrating a strong positive correlation with current sales. This relationship was substantiated by a significant coefficient and a p-value close to zero, suggesting that previous sales figures are a reliable indicator of current sales, likely capturing consumer behavior patterns or prevailing market trends more effectively than the weather variables.

In summary, these models suggest that weather conditions, particularly temperature and precipitation, have a limited direct effect on restaurant sales. The more consistent determinant of sales appears to be the historical sales data, which could be indicative of ingrained consumer habits or wider economic conditions.

5 Comprehensive examination of time series models

Understanding how sales evolve over time and predicting their future changes are critical for effective business planning. To tackle this challenge, this study employs a diverse array of time series models, each chosen for its potential to reveal different aspects of sales data. This section delves into the details of these models, exploring how they can help us interpret past sales data and forecast future trends.

Among the models chosen are both linear ones, which are grounded in well-established statistical methods, and machine learning models, which are more recent but have rapidly gained popularity for their flexibility and power in handling complex datasets. The use of machine learning in time series analysis, as highlighted by Masini et al. [8], represents a modern approach that can adapt to the non-linear patterns often found in sales data.

These models will be systematically compared to assess their strengths and weaknesses in the context of forecasting sales for the six restaurants in our study. By examining a variety of models, we aim to select the most effective approach for each restaurant's unique sales pattern, ensuring that the forecasts are as accurate and reliable as possible. To gain a deeper comprehension of the sales trends over time and to formulate forecasts, a range of models have been chosen. These models are as follows:

1. ARIMA model and its extensions: SARIMA and SARIMAX
2. Random forest
3. Prophet
4. TBATS
5. Holt-winters
6. ETS
7. Bayesian regression

5.1 ARIMA Model and its extensions

In the realm of time series analysis, the ARIMA model, along with its sophisticated extensions SARIMA and SARIMAX, are quintessential tools for unraveling and forecasting complex data dynamics. These models are renowned for their ability to parse and predict data trends through a combination of autoregressive, integrative, and moving average components, each tailored to specific characteristics of the time series.

5.1.1 ARIMA (AutoRegressive Integrated Moving Average):

The comprehensive ARIMA model [12] is denoted as ARIMA(p, d, q) and formulated as: amsmath

$$(1 - \sum_{i=1}^{p} \phi_i L^i)(1 - L)^d Y_t = (1 + \sum_{i=1}^{q} \theta_i L^i)\varepsilon_t \tag{1}$$

where:Y_t: is the time series data ε_t: is the error term L: is the lag operator **AR(p): Autoregressive Part** - It portrays the relationship between an observation and several lagged observations. The parameter 'p' represents the number of lag observations included.

$$AR(p) = (1 - \sum_{i=1}^{p} \phi_i L^i)Y_t \tag{2}$$

I(d): Integrated Part - It represents the differencing of the series to make it stationary, with 'd' being the order of differencing.

$$I(d) = (1 - L)^d Y_t \tag{3}$$

MA(q): Moving Average Part - It involves modeling the error term as a linear combination of error terms occurring contemporaneously and at various times in the past. The 'q' denotes the number of lagged forecast errors.

$$MA(q) = (1 + \sum_{i=1}^{q} \theta_i L^i)\varepsilon_t \tag{4}$$

5.1.2 Extension to ARIMAX

ARIMAX, which stands for AutoRegressive Integrated Moving Average with Exogenous Variables [11], is another powerful time series forecasting method that leverages the benefits of exogenous variables. ARIMAX extends the basic ARIMA model by incorporating external variables (X) into the forecasting process. In ARIMAX, you have the ARIMA components (p, d, q)tomodel the autocorrelation and differencing of the time series.

5.1.3 Extension to SARIMA (Seasonal ARIMA):

SARIMA [17], which stands for Seasonal AutoRegressive Integrated Moving Average, augments the ARIMA model by incorporating seasonality, a component crucial for datasets exhibiting periodic fluctuations. While the ARIMA model captures the non-seasonal patterns in time-series data, it does not inherently model seasonality.

In SARIMA, the non-seasonal part of the model is denoted as ARIMA(p, d, q), with p representing the order of the autoregressive part, d the degree of differencing, and q the order of the moving average process. The seasonal components are denoted as $(P, D, Q)_m$, where P is the seasonal autoregressive order, D is the order of seasonal differencing, Q is the seasonal moving average order, and m indicates the number of periods in each season.

The SARIMA model can be mathematically expressed as follows:

$$(1 - \textstyle\sum_{i=1}^{p} \phi_i L^i)(1 - \sum_{i=1}^{P} \Phi_i L^{i\cdot m})(1 - L)^d (1 - L^m)^D X_t = (1 + \sum_{i=1}^{q} \theta_i L^i)(1 + \sum_{i=1}^{Q} \Theta_i L^{i\cdot m})\varepsilon_t \tag{5}$$

Here, L is the lag operator, ϕ and Φ represent the non-seasonal and seasonal autoregressive parameters, θ and Θ are the non-seasonal and seasonal moving average parameters, respectively, and ε_t is the error term at time t.

5.1.4 Further Extension to SARIMAX:

Building upon SARIMA, the Seasonal AutoRegressive Integrated Moving Average with exogenous variables (SARIMAX) model further enhances the analysis by introducing external or exogenous variables, denoted by X. This extension is pivotal, especially when these exogenous factors have a significant impact on the time series. The SARIMAX model, therefore, not only captures the internal dynamics of the series through the ARIMA and seasonal components but also adjusts for the influence of external variables, providing a more comprehensive forecasting tool.

In essence, the SARIMAX model enriches the time series analysis by accommodating broader data influences, which is especially beneficial in scenarios where external factors play a critical role in shaping the series' behavior. This makes SARIMAX an invaluable model for forecasting in complex, real-world situations where multiple forces interact to influence the observed values.

5.1.5 Procedures for the Linear Models

Initially, we undertook the task of parameter identification for both ARIMA (AutoRegressive Integrated Moving Average) and SARIMA (Seasonal ARIMA) models. This involved determining the appropriate values for parameters such as autoregressive order (p), differencing order (d), moving average order (q), seasonal autoregressive order (P), seasonal differencing order (D), seasonal moving average order (Q), and the seasonal period (s). Subsequently, we proceeded with parameter estimation by fitting the chosen model to the available data. Following parameter estimation, a diagnostic check was conducted to assess the model's adequacy. Special attention was given to ensuring that the model's residuals exhibited no autocorrelation and followed a normal distribution. To accomplish this, we employed the Ljung-Box test to scrutinize residual autocorrelation, aiming for a p-value exceeding 0.05. This step was pivotal in establishing the model's robustness, affirming the null hypothesis of no autocorrelation in the residuals. In the pursuit of identifying the best models, we also utilized the auto.arima function, which allowed us to automate the model selection process. After automating the selection of various models, we carefully evaluated their performance, retaining only the most promising ones for further analysis. In cases where the diagnostic check did not meet the desired criteria, an iterative process was initiated, wherein the model specification was revised repeatedly, reiterating the aforementioned steps until satisfactory results were achieved. This meticulous procedure was meticulously followed to ensure the reliability and robustness of the model employed for predicting restaurant revenue.

5.2 Random Forest for Time Series Analysis

The Random Forest algorithm[2] stands as a paragon of ensemble learning within the realm of supervised machine learning techniques. Renowned for its versatility, it is adept at addressing both classification and regression challenges.

Fundamentally, Random Forest synthesizes the predictions of numerous decision trees, culminating in an aggregate output that typically surpasses the predictive prowess of any single constituent tree. This amalgamation leverages the strength of collective decision-making, thereby mitigating the risks of overfitting associated with individual trees and enhancing the model's generalization capabilities[10].

A particularly salient feature of the Random Forest algorithm is its intrinsic mechanism for feature selection. It inherently evaluates and ascertains the relative importance of various predictors, thereby facilitating a focused model that capitalizes on the most influential variables. In practical implementation, especially within the R programming environment, this aspect can be harnessed to refine the model's accuracy and efficiency in uncovering the underlying patterns within the temporal data.

Thus, the Random Forest algorithm emerges as an efficacious tool in time series analysis, adept at navigating the intricacies of temporal data to yield robust, reliable forecasts.

5.3 Prophet Time Series Forecasting Model

Prophet, conceptualized by Facebook Research [15], stands as a contemporary approach to univariate time series forecasting. This model is distinguished for its user-friendly implementation and robustness, making it adept at analyzing data characterized by trends and seasonal variations. Notably, Prophet effectively manages missing data and outliers, underscoring its practical utility in diverse forecasting scenarios.

The core of the Prophet model is a decomposable time series framework, encompassing three primary components: trend ($g(t)$), seasonality ($s(t)$), and holidays ($h(t)$). The aggregate forecast

$y(t)$ is formulated as the sum of these components, supplemented by an error term ε_t:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t \tag{6}$$

The trend component $g(t)$ is engineered to capture non-periodic changes within the dataset. Depending on the data's characteristics, $g(t)$ can represent a non-linear trend modeled using logistic growth or a piecewise linear trend for linear data progressions. The logistic growth model is expressed as:

$$g(t) = \frac{C(t)}{1 + \exp(-(k + a(t)^\top \delta)(t - (m + a(t)^\top \gamma)))} \tag{7}$$

Here, $C(t)$ is the carrying capacity, k is the growth rate, δ and γ are adjustment parameters, and $a(t)$ is a vector of seasonality adjustments.

The seasonality component $s(t)$ incorporates periodic changes, such as weekly or annual cycles, using Fourier series. This method endows Prophet with the flexibility to model complex seasonal behaviors. The seasonal effect is calculated as a sum of sine and cosine terms with estimated coefficients β:

$$s(t) = \sum_{n=1}^{N} (a_n \cos(\frac{2\pi nt}{P}) + b_n \sin(\frac{2\pi nt}{P})) \tag{8}$$

In this expression, P represents the period (e.g., 365.25 for yearly seasonality), and N is the number of Fourier terms used.

The holiday component $h(t)$ addresses the influence of holidays and events, which are modeled using an indicator function $Z(t)$ that alters the forecast in accordance with the occurrence of these events:

$$h(t) = Z(t)k \tag{9}$$

Each holiday is represented by $Z(t)$, with a corresponding change k in the forecast.

The Prophet models ability to decompose time series data into these interpretable components, paired with its inherent flexibility, renders it a powerful tool in the domain of forecasting. It excels in scenarios where seasonal patterns and irregular events play a significant role, offering insights and predictive accuracy that align closely with real-world trends.

5.4 TBATS Time Series Forecasting Model

The BATS and TBATS[4] models are advanced forecasting techniques that have been developed to handle time series data with complex seasonal patterns. The reduced form of these models is represented by the equation:

$$\phi_p(L)\eta(L)y(\omega)_t = \theta_q(L)\delta(L)\varepsilon_t \tag{10}$$

In this equation: - $\phi_p(L)$ and $\theta_q(L)$ are the autoregressive and moving average polynomials of orders p and q respectively. - $\eta(L)$ represents the component capturing seasonality and trend. - $y(\omega)_t$ is the transformed time series at time t. - $\delta(L)$ is a polynomial that encapsulates additional seasonal structures. - ε_t is the error term.

The expanded form of the TBATS model, particularly for the case with non-stationary growth, further explicates the $\eta(L)$ and $\delta(L)$ components:

$$\eta(L) = (1-L)(1-\phi L)^T \prod_{i=1}^{T} \prod_{j=1}^{k_i} (1 - 2\cos\lambda_j^{(i)} L + L^2) \tag{11}$$

$$\begin{aligned}
\delta(L) = & \left[L^2\phi(1-\alpha) + L(\alpha + \phi\beta - \phi - 1) + 1\right] \times \prod_{i=1}^{T} \prod_{j=1}^{k_i} (1 - 2\cos\lambda_j^{(i)} L + L^2) \\
& + (1-L)(1-\phi L) \times \sum_{i=1}^{T} \sum_{j=1}^{k_i} \left[\prod_{\tilde{i}=1, \tilde{i}\neq i}^{T} \prod_{\tilde{j}=1, \tilde{j}\neq j}^{k_{\tilde{i}}} (1 - 2\cos\lambda_{\tilde{j}}^{(\tilde{i})} L + L^2) \right] \\
& \times \left[(\cos\lambda_j^{(i)} \gamma_{1i} + \sin\lambda_j^{(i)} \gamma_{2i}) L^2 - \gamma_{1i} L^3 \right] \\
& + (1-L)(1-\phi L)L \times \prod_{i=1}^{T} \prod_{j=1}^{k_i} (1 - 2\cos\lambda_j^{(i)} L + L^2) \times \sum_{i=1}^{T} k_i \gamma_{1i}
\end{aligned} \tag{12}$$

In these expanded expressions: - $\eta(L)$ includes the differencing and seasonal differencing terms, reflecting non-stationary and seasonal components. - $\delta(L)$ comprises a series of trigonometric expressions that model the complex seasonal patterns within the time series. The trigonometric terms, involving $\cos\lambda_j^{(i)}$ and $\sin\lambda_j^{(i)}$, allow the model to adapt to the seasonality of different frequencies, denoted by $\lambda_j^{(i)}$, which are specific to each seasonal component i and its harmonics j.

The comprehensive structure of the TBATS model, as encapsulated by these equations, enables it to adeptly handle various types of seasonalities and trends in time series data. This capability makes it particularly suited for forecasting scenarios where traditional models may struggle to capture the underlying patterns in the data.

5.5 ETS (Error-Trend-Seasonality) Model

The ETS (Error-Trend-Seasonality) model[5] stands as a prominent methodology in the realm of time series forecasting. This model adeptly captures the inherent characteristics of data through its core components: error, trend, and seasonality. Central to ETS is the notion of decomposing time series data into these fundamental elements, each representing a distinct aspect of the data's behavior.

The simplest incarnation of the ETS model, often referred to as simple exponential smoothing, is conceptualized as an (A, N, N) model. This configuration entails additive errors, the absence of a trend component, and no seasonality. It can be represented in a state space form as follows:

$$y_t = l_{t-1} + e_t \tag{13}$$

$$l_t = l_{t-1} + \alpha e_t \tag{14}$$

Here, y_t represents the observed data at time t, l_t denotes the level of the series, and e_t is the error term. The parameter α is the smoothing constant for the level.

In ETS models, the error component encapsulates random fluctuations that are not explained by the trend or seasonality. The trend component, on the other hand, captures the long-term direction in which the data is moving, be it an upward, downward, or stable trend. The seasonality component is crucial for data exhibiting periodic patterns at regular intervals, such as daily, weekly, or monthly cycles.

The flexibility of the ETS model lies in its ability to adapt to different data characteristics. It offers both additive and multiplicative options for combining the components, catering to datasets with constant or proportional fluctuations. Additionally, the model can incorporate a damped trend, moderating the influence of the trend over time. The choice of error typesGaussian or exponential, for instancefurther tailors the model to the specific error distribution observed in the data.

This model provides a comprehensive framework for analyzing time series data, accommodating a wide range of patterns and allowing for precise forecasting based on the underlying structural components of the dataset.

5.6 Holt-Winters forecasting technique

Holt-Winters [6]is a widely used technique for predicting time series data. Despite its age, it remains valuable for tasks like anomaly detection and capacity planning. It relies on exponential smoothing to capture historical data patterns and use them to predict typical values for the present and future. Think of it as a more advanced version of the simple moving average method.

In Holt-Winters, the three key aspects of time series behavioraverage, trend, and seasonalityare expressed using three types of exponential smoothing:

1. **Single exponential smoothing:** This is suitable for predicting data without a clear trend or seasonal pattern, where the data's level might change over time.

2. **Double exponential smoothing:** This is used for predicting data with a noticeable trend.

3. **Triple exponential smoothing:** This method is applied when there's a trend and/or seasonality in the data.

These components are combined to make predictions, making Holt-Winters suitable for non-stationary time series.

The following are the formulas for both the additive and multiplicative forms of the model: **Additive form:**

Forecast:

$$\hat{y}_{t+h|t} = l_t + hb_t + s_{t+h-m(k+1)} \tag{15}$$

Level:

$$l_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(l_{t-1} + b_{t-1}) \tag{16}$$

Trend:

$$b_t = \beta \cdot (l_t - l_{t-1}) + (1-\beta) \cdot b_{t-1} \tag{17}$$

Seasonality:

$$s_t = \gamma \cdot (y_t - l_{t-1} - b_{t-1}) + (1 - \gamma) \cdot s_{t-m} \tag{18}$$

Multiplicative form:

Level:

$$l_t = \alpha \cdot \frac{y_t}{s_{t-m}} + (1 - \alpha) \cdot (l_{t-1} + b_{t-1}) \tag{19}$$

Trend:

$$b_t = \beta \cdot (l_t - l_{t-1}) + (1 - \beta) \cdot b_{t-1} \tag{20}$$

Seasonality:

$$s_t = \gamma \cdot \frac{y_t}{l_{t-1} + b_{t-1}} + (1 - \gamma) \cdot s_{t-m} \tag{21}$$

In the additive form of the Holt-Winters model, the forecasted value is a combination of the level, trend, and seasonality components.The level represents the general average, the trend captures any upward or downward movement, and the seasonality accounts for regular patterns. In the multiplicative form of the Holt-Winters model, the forecasted value is calculated by multiplying the level, trend, and seasonality components. The level represents the general average, the trend captures the direction and rate of change, and the seasonality accounts for regular patterns. The additive method is preferable when seasonal variations remain relatively constant, while the multiplicative method is suitable when seasonal variations change proportionally with the level of the series.

5.7 Bayesian Regression

Bayesian regression offers a robust approach to modeling time series data, leveraging the principles of Bayesian statistics. This method stands out for its ability to incorporate prior knowledge or beliefs into the analysis, which is particularly useful in time series forecasting where historical data can inform about future trends.

In Bayesian regression, parameters are treated as random variables with specified distributions, reflecting prior beliefs. As new data becomes available, these beliefs are updated using Bayes' Theorem, leading to posterior distributions of the parameters. This iterative process of updating beliefs with new evidence is a fundamental strength of Bayesian methods.

A key advantage of Bayesian regression in time series analysis is its capacity to handle uncertainty and incorporate complex hierarchical structures. It allows for more nuanced modeling of time-dependent patterns, such as trends and seasonality, and provides a probabilistic framework for prediction, offering not just point estimates but entire distributions of possible outcomes.

Bayesian methods are particularly powerful in situations with limited data, where incorporating prior knowledge can significantly enhance the model's accuracy and reliability. The flexibility and depth of Bayesian regression make it a valuable tool for sophisticated time series forecasting.

6 Methodological aspects

6.1 Enhanced forecasting model evaluation via rolling origin

The evaluation of forecasting models is pivotal in time series analysis, with the hold-out method serving as the foundational approach. This conventional method partitions the dataset into two segments: the training set for model development and the testing set for validation. The primary measure of model efficacy is its performance on this fixed testing set. While this method has provided a baseline for model evaluation, it is limited by its static nature. It does not account for potential changes in data patterns over time, which can result in an evaluation that does not fully reflect the model's capacity to adapt to new information.

The rolling origin evaluation method offers a solution to these limitations by introducing a dynamic and iterative approach to forecasting. This technique, as opposed to a static hold-out sample, continuously updates the forecast originshifting the point at which the model begins making predictions forward with each new data point. This method is particularly advantageous as it mitigates the risk of overfitting to a specific period's data characteristics and ensures that the model's performance is tested against various market conditions and operational scenarios.

Figure 23 encapsulates the concept of rolling origin evaluation. It portrays a sequence of forecast iterations, where the in-sample data (training set) is used to estimate the model, marked by the white cells, and the light grey cells represent the out-of-sample forecasts produced at each iteration. This visualization, derived from the work of Svetunkov and Petropoulos (2018), illustrates how the model is re-estimated and its predictions recalibrated as the forecast origin progresses through the dataset.

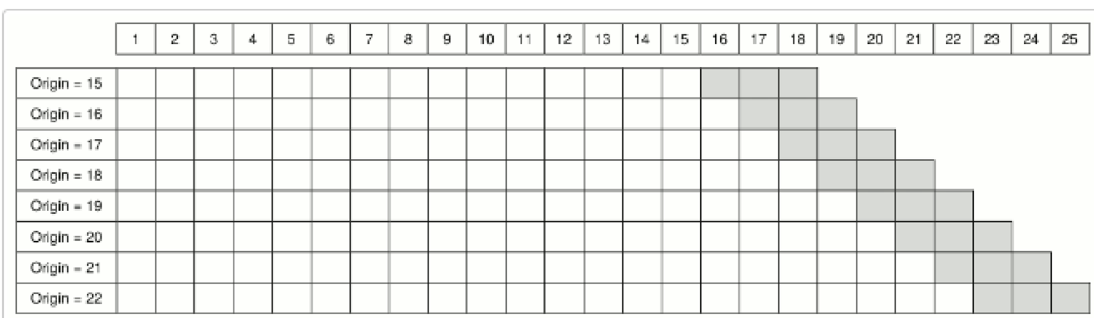

Figure 23: Rolling Origin with fixed train size

In our study, we enhance the robustness of our evaluation by implementing rolling origin with forecast horizons of 1, 7, 14, 30, and 60 days with 20 origins. This range of horizons allows us to thoroughly investigate the model's predictive accuracy across different temporal scalesfrom short-term to long-term forecasts. By employing this technique, we observe not only the general forecast error but also the variability and consistency of the model's performance over time. It provides insights into how the model responds to fluctuations in data trends and the reliability of its forecasts in the face of evolving market dynamics. The rolling origin method, therefore, presents a more detailed and rigorous assessment of forecast accuracy than the traditional hold-out method.

The comprehensive error metrics obtained from this approach offer a granular understanding of the forecasting model's behavior under varied conditions, which is crucial for businesses that rely on precise demand forecasting, such as the restaurant industry. By leveraging the rolling origin technique, we ensure that our model's validation is as exhaustive and indicative of real-world performance as possible[13][14].

6.2 Evaluation criteria

In the analysis of time series,it is crucial to use reliable metrics[7]. Two such metrics are Root Mean Squared Error (RMSE) and Mean absolute percentage error (MAPE). To evaluate our models we will use those 2 metrics.

Root Mean Squared Error (RMSE) is a widely used measure in statistics to quantify the difference between values predicted by a model and the values actually observed. This metric offers a clear idea of how much error is involved in the forecasts, on average. RMSE is computed as the square root of the average squared differences between the forecasted values and the actual values. The formula for RMSE is:

$$RMSE_h = \sqrt{\frac{1}{T-m-1}\sum_{t=m}^{T-1}(y_{t+h} - \hat{y}_{t+h|t})^2} \tag{22}$$

In this formula, y_{t+h} represents the actual observed values at time $t+h$, and $\hat{y}_{t+h|t}$ represents the forecasted values at the same point. The term T denotes the total number of observations, and m represents the starting point of the forecast period. The subscript h indicates the specific time horizon of the forecast.

Mean Absolute Percentage Error (MAPE), on the other hand, is a measure of prediction accuracy in a forecasting model. It expresses the size of error as a percentage, which makes it easy to interpret and particularly useful when comparing the accuracy of different forecasting models. MAPE is advantageous because it is scale-independent, meaning it can be used to compare forecasts across data sets with different scales. The formula for MAPE is:

$$MAPE_h = \frac{1}{T-m-1}\sum_{t=m}^{T-1}\left|\frac{y_{t+h} - \hat{y}_{t+h|t}}{y_{t+h}}\right| \tag{23}$$

In this equation, like in RMSE, y_{t+h} and $\hat{y}_{t+h|t}$ are the actual and forecasted values, respectively. The absolute value of the percentage difference between these values is averaged over all forecasted points.

Both RMSE and MAPE are essential for evaluating the accuracy of time series forecasts. They provide valuable insights into the performance of predictive models, helping analysts and decision-makers understand the effectiveness of their forecasting strategies in real-world scenarios.

6.3 Models evaluation procedure

In our analysis, we've extensively employed linear models for time series analysis, adhering to the Box-Jenkins methodology, a well-established approach in statistical analysis [1], [16]. This method, renowned for its efficacy in time series forecasting, involves several critical steps.

Initially, the 'identification' phase is paramount. Here, we closely examine the time series data to determine the most suitable model. This involves an in-depth analysis of Autocorrelation Function

(ACF) and Partial Autocorrelation Function (PACF) graphs. These tools are instrumental in identifying the order of Autoregressive (AR) and Moving Average (MA) components in the ARIMA model. The meticulous analysis ensures that the selected model aptly captures the underlying patterns and characteristics of the time series data.

Upon identifying a plausible model, we proceed to the 'estimation' phase. This critical step involves the use of maximum likelihood estimation methods to accurately determine the model's parameters. The careful calibration of these parameters is crucial for the model's ability to predict future values reliably.

Subsequent to model estimation, we conduct a rigorous 'diagnostic checking' phase. This phase is crucial for validating the model's adequacy and involves a thorough examination of the model's residuals. Ensuring that the residuals are independent and exhibit a normal distribution is vital for the model's validity. This necessitates the implementation of the Ljung-Box test [3], which scrutinizes the randomness and independence of the residuals. Models that fail to meet these stringent criteria are subjected to further refinement, iterating through the process until a robust and reliable model is achieved.

The culmination of this meticulous process is the 'forecasting' phase where the finalized model is employed to predict future values of the time series. These predictions are grounded in the historical data and the reliability of the model, established through the previous phases.

The application of this methodology in our research has enabled us to develop models that are not only statistically sound but also highly effective in forecasting. The iterative nature of this process, coupled with stringent diagnostic checks, ensures that our models are both robust and capable of capturing the intricate dynamics of time series data.

7 Predictive Analysis

This part of the study is where we try to answer to two of the initial research questions. The aim of the predictive models, is to first of all understand how would have been the earnings during the lockdown, during the period that goes from the 9 March 2020 to the 5 of May 2020 and furthermore for a missing month after the registered data in the future trough all the 6 restaurants. The models that we are going to apply are:

- ARIMA and its extensions
- Prophet
- Random Forest
- TBATS
- HoltWinters
- ETS
- Bayesian regression

7.1 Predictions for Covid-19

In order to conduct a more rigorous analysis aiming at precise forecasting of potential restaurant earnings, the time series under examination encompasses the period spanning from September 1, 2018, to February 19, 2020. series dataset, devoid of speculative influences or extraneous variables, thereby rendering it suitable for the training of predictive models and insulating the data from the perturbing effects of the subsequent viral emergence.

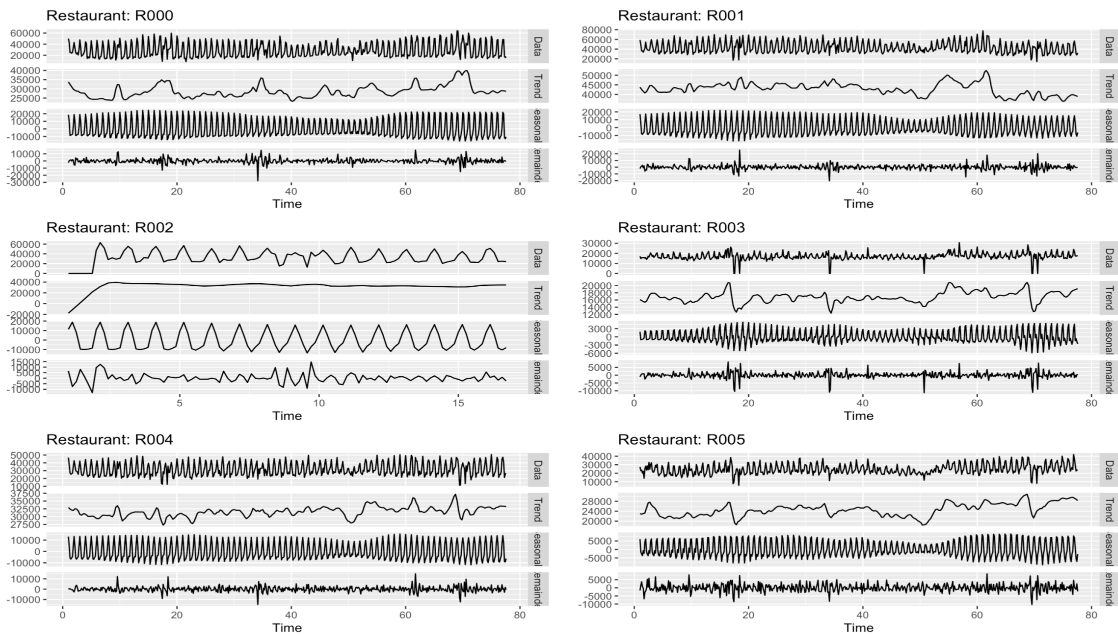

Figure 24: Figure 20: Decomposition of the time series Pre-Covid

We proceeded with the decomposition of the historical series from September 1, 2018, to February 19, 2020, considering the data related to sales. This operation was necessary since the historical series have different patterns inside them and thanks to their use of particular models it is possible to divide the time series in the various components that make it up. In case examined, the components turn out to be trends, weekly seasonality and residuals. So this time series shows clear signs of seasonality. During this time interval, it is possible to observe significant shrinkages followed by a series of broadenings of the data series. This seasonal structure suggests a clear trend in the behavior of the data over these years. Upward spikes may reflect periods of increased activity, seasonal events or fluctuations in demand. On the other hand, downward peaks in the trend may be associated with periods of low activity or less favorable economic conditions such as upward peaks in economic recoveries and favorable events.

Reviewing the sales trends of the five restaurants, we observe that Restaurants R000 and R005 showcase similar stable trends with consistent periodic fluctuations, hinting at a regular customer base or effective sales strategies. Restaurants R001 and R004 both display high volatility in their sales patterns, with sharp peaks suggesting the influence of intermittent, high-impact events. Restaurant R002 stands out with its clear cyclical behavior, indicating a strong weekly sales rhythm, while Restaurant R003 presents a more stable demand with less pronounced fluctuations compared to the others

7.2 Arimax and Sarimax with Fourier Series

In this analysis, we aimed to create a model for predicting a time series dataset by employing an ARIMAX model with Fourier terms. The main goal was to effectively capture the underlying patterns and seasonality within the data for accurate forecasting. In line with the principle of Occams Razor, we initially explored models that considered both weekly and yearly seasonality terms. However, after careful considerations, we opted for a simpler approach by selecting the model with only weekly seasonality. This choice aligns with the concept of preferring simpler solutions when multiple equally valid options are available. To account for the cyclic patterns present in the time series data, Fourier terms were introduced. These terms allow the model to capture cyclic patterns, such as weekly and yearly seasonality, without introducing an excessive number of individual variables. By leveraging Fourier terms to represent seasonality, we achieved a more concise yet effective model. The ARIMA model, which combines AutoRegressive (AR) and Moving Average (MA) components, was utilized to build the model. Additionally, Fourier terms were included as exogenous variables to help the model capture both the cyclic patterns and the autoregressive and moving average dynamics. Our model underwent training using the training dataset, and its in-sample predictions were evaluated using the Root Mean Square Error (RMSE) metric. This provided a measure of how accurately the model reproduced the observed data points. Parameters and performance statistics of the model were summarized for a comprehensive understanding. Subsequently, the model was used to generate forecasts for the test dataset. Since this is the first model presented the residual analysis has been performed on all restaurants.

Restaurant	Model	RMSE	MAPE
R000	ARIMA(1,0,0)(1,0,0)[7]	7501.44	18.70
R001	ARIMA(1,1,1)(1,0,0)[7]	7907.73	19.57
R002	ARIMA(0,0,0)	3191.25	8.12
R003	ARIMA(1,1,0)(0,0,1)[7]	3376.73	8.43
R004	ARIMA(3,1,1)	4116.19	9.45
R005	ARIMA(2,1,2)(0,0,1)[7]	4088.80	9.54

Table 1: Summary of ARIMA models with Fourier transformations performances

For Restaurant R000, employing the ARIMA(1,0,0)(1,0,0)[7] model resulted in a RMSE of 7501.44, indicating its predictive accuracy, Restaurant R002 demonstrated notable performance with an ARIMA(0,0,0) model, achieving a low RMSE of 3191.25. The ARIMA(1,1,0)(0,0,1)[7] model for Restaurant R003 yielded a RMSE of 3376.73, showcasing its effectiveness in forecasting. Restaurants R004 and R005, utilizing ARIMA(3,1,1) and ARIMA(2,1,2)(0,0,1)[7] models, respectively, displayed RMSE values of 4116.19 and 4088.80, indicating comparable predictive performance across these establishments.

Restaurant	Daily Average Loss ()	Total Loss ()
R000	13,943	976,000
R001	27,005	1,890,300
R002	18,380	1,286,600
R003	16,950	1,185,500
R004	27,343	1,914,000
R005	21,433	1,500,300

Table 2: Summary of Restaurant Losses

7.3 SARIMAX with daily dummies

In this case we incorporated a days dummy variable into the ARIMA models for each restaurant (ARIMA(2,1,3) model for R004 as example) to effectively capture recurring seasonality patterns within our restaurants revenue data. This variable is imperative for improving the models capacity to discern and adapt to variations associated with specific days of the week as espected from the EDA.

Restaurant	Model	RMSE	MAPE
R000	ARIMA(3,0,2)	7563.87	19.89
R001	ARIMA(4,0,1)	8528.76	20.11
R002	ARIMA(1,0,4)	3695.36	8.45
R003	ARIMA(0,1,2)	3499.84	8.23
R004	ARIMA(2,1,3)	4160.54	9.02
R005	ARIMA(2,1,3)	4208.97	10.53

Table 3: Summary of ARIMA models performances

Notably, for Restaurant R002, the ARIMA(1,0,4) model demonstrated a relatively low RMSE of 3695.36, showcasing its effectiveness in predicting the gross amount. The ARIMA(0,1,2) model exhibited a commendable performance for Restaurant R003, yielding a RMSE of 3499.84, indicative of its accuracy in forecasting the target variable. Restaurants R004 and R005 both utilized the ARIMA(2,1,3) model, resulting in RMSE values of 4160.54 and 4208.97, respectively, suggesting a comparable predictive performance across these establishments.

As we trained a model for each restaurant here's the results of the predicted losses during lockdown:

Restaurant	Average	Total
R000	27908.08	1953565.70
R001	45419.11	3179337.60
R002	29193.31	2043531.63
R003	18590.24	1301317.07
R004	32749.50	2292464.85
R005	28188.70	1973208.66

Table 4: Average and total estimated loss for each restaurant

7.4 Random Forest

For the Random Forest analysis, a deliberate method was employed for variable selection using the importance function. Variables boasting an importance score above zero were judiciously retained, as elucidated by the corresponding visual representations. Such a meticulous approach notably enhanced the Random Forest models performance.

Previously, the models predictions oscillated between values of 3k and 5k, and there were discernible patterns in the residuals. However, with the integration of more relevant variables and a particular emphasis on the inclusion of lag variables, there was a marked improvement in the models stability, ushering in more consistent predictions. This refinement not only improved the overall reliability of the results but also led to an observation where the residuals were devoid of autocorrelation, underscoring the paramount role played by the lag variables. The integration of lag variables as pivotal features in models, especially in the likes of the Random Forest, unfolds a myriad of advantages. They adeptly model the autocorrelation inher- ent in time series data, capturing the innate correlation a series has with its preceding values.

For instance, the presence of a one-step lag encapsulates the relationship between a value and its immediate predecessor. Moreover, for datasets with daily entries, the inclusion of a seventh lag might reveal weekly effects, like the recurrent peaks in weekend sales. While conventional time series models, such as ARIMA, are designed to use lag variables, the Random Forest stands out in its ability to weave these lags with other explanatory variables, unearthing non-linear patterns and intricate interactions. This selective integration of lag variables, especially when underpinned by domain-specific knowledge and metrics like partial autocorrelation, equips the Random Forest to unravel temporal intricacies without wading into over-complexity, thus averting potential overfitting.

Furthermore, in scenarios where the impact of a particular lag morphs over a duration, for example, where the preceding days effect varies across seasons, the Random Forest emerges as a robust tool. Its inherent design enables it to discern and adapt to these non-linear dynamics, providing an edge over traditional linear models. This flexibility is further enhanced by the models ability to perceive and decipher interactions between lag variables and other determinants, presenting a holistic understanding of the temporal intricacies involved.

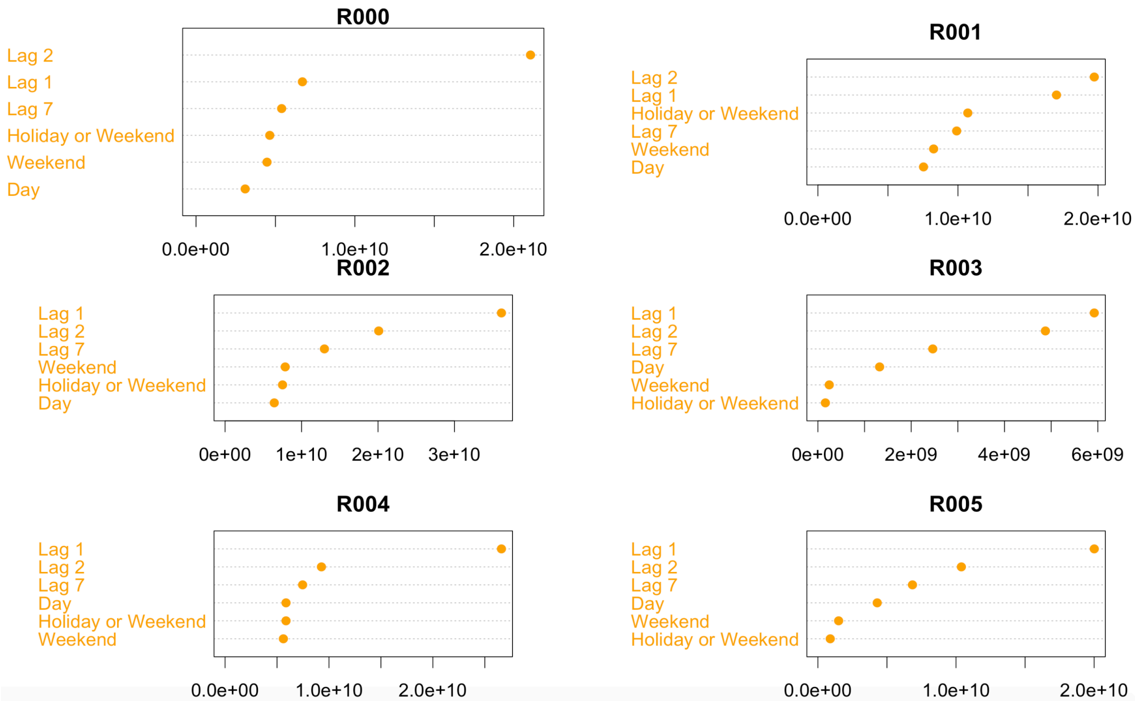

Figure 25: Feature importance in Random Forest with Increase in Node Purity as metric

The examination of feature importance in random forest models applied to diverse restaurant time series data reveals common trends and notable distinctions. Features such as "Day," "Weekend," and "Holiday or Weekend" consistently emerge as influential across various establishments, indicating shared predictive factors. However, the importance of temporal patterns, exemplified by lag features (`lag_1`, `lag_2`, `lag_7`), varies noticeably. For instance, in Restaurant R002, `lag_1` and `lag_2` carry exceptionally high importance, suggesting a pronounced reliance on short-term historical data. In contrast, Restaurant R004 prioritizes `lag_2` more than other lags, showcasing unique predictive patterns.

Restaurant	RMSE	MAPE (%)
R000	7808.21	18.64
R001	7255.13	18.45
R002	6605.05	14.02
R003	3392.93	10.08
R004	4049.47	11.77
R005	5077.39	17.09

Table 5: Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) for each restaurant Random Forest

The table25 presents the results of the predictions using random forest model for the six restaurants. Notably, Restaurants R003 and R002 demonstrate the lowest RMSE values at 3392.93 and 6605.05, respectively, indicating more accurate predictions. R003 also stands out with the lowest MAPE at 10.08%, reflecting a smaller percentage error. Conversely, R000 and R005 exhibit higher RMSE and MAPE values, suggesting a larger deviation between predicted and actual values and a higher proportional error. As we trained a model for each restaurant the results of the predicted losses during lockdown can be observed in fig.7:

Table 6: Restaurant Loss Data

Restaurant	**Daily Average Loss ()**	**Total Loss ()**
R000	25793.72	1805560.15
R001	43576.05	3050323.78
R002	19717.23	1380205.85
R003	13294.60	930621.81
R004	23561.07	1649274.86
R005	18596.05	1301723.52

Table 7: Average and total estimated loss for each restaurant

7.5 Prophet

The adoption of the Prophet forecasting model is grounded in its exceptional ability to effectively address the complexities of multiple seasonalities, a facet often overlooked in alternative forecasting tools. This nuanced approach to seasonality ensures the comprehensive capture of overt and subtle temporal patterns, resulting in forecasts characterized by precision and faithful representation of underlying temporal dynamics. In essence, our selection of Prophet was driven by its unparalleled proficiency in comprehending and modeling the inherent rhythms and cycles within our dataset. Additionally, the incorporation of a dummy daily variable has proven instrumental in enhancing the models data handling capabilities. To assess how well the model fits the data.

Restaurant	**RMSE**	**MAPE (%)**
R000	6853.52	19.93
R001	12694.54	23.36
R002	3448.27	8.07
R003	3807.94	14.46
R004	4116.89	9.93
R005	4092.67	10.19

Table 8: Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) for each restaurant Prophet

The provided table outlines the prediction results for various restaurants using the Prophet forecasting model. Notably, Restaurant R002 stands out with the lowest RMSE at 3448.27, indicating precise predictions, and a remarkably low MAPE of 8.07%, reflecting a small percentage error. On the other hand, R001 exhibits a relatively high RMSE of 12694.54 and a considerable MAPE of 23.36%, suggesting less accurate predictions and a higher proportional error. The table offers insights into the varying performance of the Prophet model across different restaurants, highlighting its effectiveness in some cases and potential limitations in others. Looking the estimates shown by the model, the revenue lost during the closing of the lookdown is showed in the table below.

Table 9: Restaurant loss data

Restaurant	Daily Average Loss ()	Total Loss ()
R000	27997.01	1959790.70
R001	43972.66	3078086.01
R002	35149.86	2460490.01
R003	16936.86	1185579.91
R004	31359.04	2195132.50
R005	23913.99	1673979.50

Table 10: Average and total estimated loss for each restaurant

7.6 Holt-Winters

We employed the Holt-Winters Exponential Smoothing method for our time series forecasting. Initially, we plotted the training data to get an overview of its characteristics. We then proceeded with fitting the Holt- Winters model, selecting a multiplicative seasonal component. This decision was driven by warnings that arose when applying an additive seasonality model, indicating a potential misfit. Hence, the multiplicative model was considered a better fit for the datas nature. Upon fitting the model to our training data, the method optimized the smoothing parameters to achieve the best fit. The resulting parameter values as example for R004 were as follows:

- **Alpha (level parameter):** $\alpha = 0.0586$
- **Beta (trend parameter):** $\beta = 0.0211$
- **Gamma (seasonal parameter):** $\gamma = 0.1593$

These parameters play a crucial role in determining how the model responds to changes in the level, trend, and seasonality of the data. The relatively low values of alpha and beta suggest that the model assigns more significance to past observations when adjusting for changes in the level and trend, respectively. The gamma value of 0.1593 indicates a moderate weight assigned to past seasonal patterns.

Restaurant	RMSE	MAPE (%)
R000	7120.68	20.87
R001	12345.67	33.21
R002	3680.11	8.62
R003	3958.43	19.21
R004	4256.78	12.35
R005	4173.26	13.78

Table 11: RMSE and MAPE for each restaurant Holt-Winters

The table 25 presents the prediction results for various restaurants using the Holt-Winters forecasting model. Notably, Restaurant R002 demonstrates the lowest RMSE at 3680.11, indicating precise predictions, and a remarkably low MAPE of 8.62%, reflecting a small percentage error. Conversely, R001 exhibits a relatively high RMSE of 12345.67 and a considerable MAPE of 33.21%, suggesting less accurate predictions and a higher proportional error. These results provide insights into the varying performance of the Holt-Winters model across different restaurants, emphasizing its effectiveness in specific cases and potential challenges in others. In the graph below we can see how the gross amount might have been in the covid period using the model previously trained.

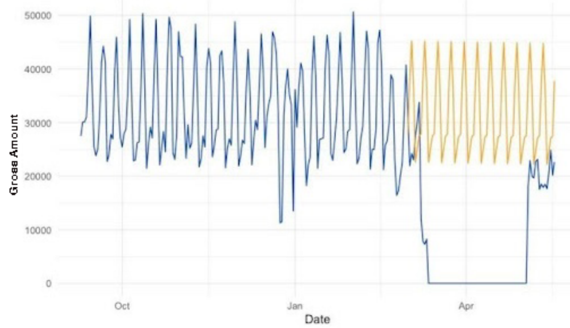

Figure 26: Forecasting visualization for the COVID period in Restaurant R004 using Holt-Winters method.

Looking the estimates shown by the model, the revenue lost during the closing of the lookdown is showed in the table 25 .

Table 12: Restaurant loss data

Restaurant	Daily Average Loss ()	Total Loss ()
R000	28765.43	2013554.92
R001	44532.11	3117245.78
R002	36218.97	2536320.88
R003	17689.24	1238256.07
R004	32658.12	2287073.21
R005	24659.83	1725734.91

Table 13: Average and total estimated loss for each restaurant

7.7 ETS

To further validate our forecasting approach, we explored the Error, Trend, Seasonality (ETS) model, which offers another method for time series forecasting. Given the identifiable systematic and seasonal patterns in the data, we selected the ETS model with an additive error, no trend, and additive seasonality (ETS(A,N,A)) as the most suitable choice. Upon applying the ETS(A,N,A) model to the training dataset, the method determined the optimal smoothing parameters. These parameters provide insights into how the model handles the level, trend, and seasonality components of the data:

- **Alpha (level parameter):** $\alpha = 0.0444$
- **Gamma (seasonal parameter):** $\gamma \approx 0$ (very close to zero, approximately 1×10^{-4})

As example for R004 the alpha value suggests that the model places significant weight on past observations when adjusting the current level of the series. This characteristic imparts stability to the models predictions. Conversely, the near-zero gamma value indicates that the model minimizes its reliance on recent seasonal changes, instead leaning heavily on the established seasonal pattern.

Restaurant	RMSE	MAPE (%)
R000	7807.77	31.17
R001	12157.04	32.47
R002	3011.82	7.64
R003	3284.72	15.19
R004	4115.22	11.90
R005	4237.75	15.09

Table 14: RMSE and MAPE for each restaurant using ETS

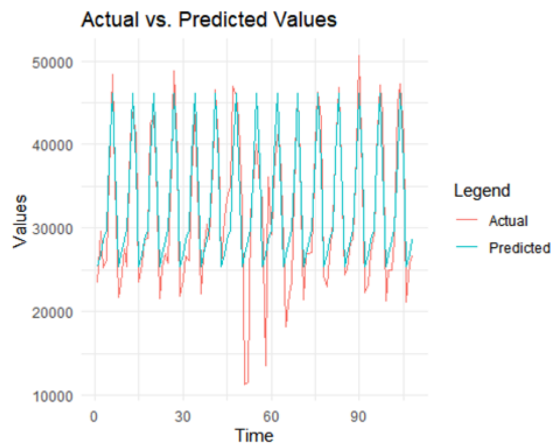

Figure 27: Forecasting visualization for covid period on R004

The table presents the prediction results for various restaurants using the ETS (Error, Trend, and Seasonality) forecasting model, with metrics reported in terms of Root mean square error and mean absolute percentage error. Notably, Restaurant R002 demonstrates the lowest RMSE at 3011.82, indicating precise predictions, and a remarkably low MAPE of 7.64%, reflecting a small percentage error. Conversely, R000 and R001 exhibit relatively high RMSE values of 7807.77 and 12157.04, respectively, along with corresponding MAPE values of 31.17% and 32.47%. These results offer insights into the varying performance of the ETS model across different restaurants, emphasizing its effectiveness in specific cases and potential challenges in others.

To provide a visual representation, we propose as example on R004 2 plots that showcases both actual and predicted values over time. This visualizations enables a clear comparison of prediction value is against the real data. As can be seen from the second forecast graphs, the expected values during the covid are in the range of 20000 and 45000.

Looking the estimates shown by the model, the revenue lost during the closing of the lookdown is showed in the table below.

Table 15: Restaurant loss data

Restaurant	Daily Average Loss ()	Total Loss ()
R000	35307.87	2471551.16
R001	49635.62	3474493.27
R002	33105.70	2317398.82
R003	18983.95	1328876.66
R004	33067.75	2314742.54
R005	28215.99	1975119.14

Table 16: Average and total estimated loss for each restaurant

7.8 TBATS

The TBATS method was employed for time series analysis, chosen for its capability to effectively handle multiple seasonalities. It is crucial to emphasize that this modeling approach does not facilitate the inclusion of external regressors, such as dummy variables. This limitation arises from the inherent capacity of the TBATS method to adeptly manage multiple seasonal patterns within the time series data.

A noteworthy parameter in the model configuration is the fixed value of alpha (α) set at 0.024. This parameter plays a pivotal role in assigning heightened significance to the seasonal component inherent in the time series data.

Restaurant	RMSE	MAPE (%)
R000	7127.59	18.27
R001	12465.53	33.91
R002	2826.97	6.75
R003	3235.32	14.84
R004	4174.68	11.99
R005	4139.44	14.59

Table 17: Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) for each restaurant TBATS

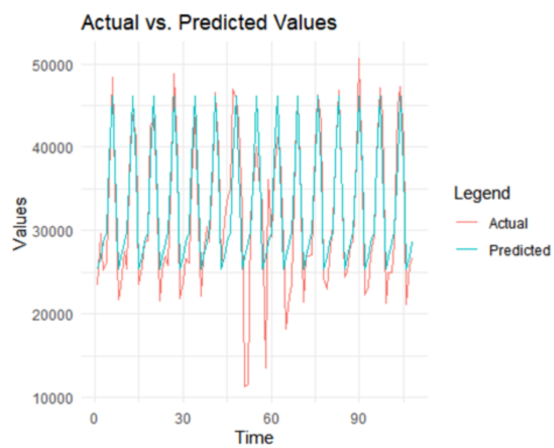

Figure 28: TBATS Actual vs Predicted on test set on R004

Notably, Restaurant R002 exhibits the lowest RMSE at 2826.97, indicating precise predictions, and a remarkably low MAPE of 6.75%, reflecting a small percentage error. Conversely, R001 displays a relatively high RMSE of 12465.53 and a considerable MAPE of 33.91%, suggesting less accurate predictions and a higher proportional error. These results offer insights into the varying performance of the TBATS model across different restaurants, emphasizing its effectiveness in specific cases and potential challenges in others.

Looking the estimates shown by the models, the revenue lost during the closing of the lookdown is showed in the table 25.

Restaurant	Daily Average Loss ()	Total Loss ()
R000	29172.63	2042084.30
R001	50021.02	3501471.17
R002	32688.94	2288226.14
R003	18734.74	1311432.03
R004	32869.95	2300896.31
R005	27927.91	1954953.37

Table 18: Average and total estimated loss for each restaurant TBATS

7.9 Bayesian linear regression model

The initial analysis utilized a basic Bayesian model, which primarily aimed to lay down a foundational predictive performance. In this rudimentary model, emphasis was given to the "Day" variable, representing the day of the week. This inclusion was strategic, given the variables efficacy in enhancing model performance, combined with a curiosity to gauge the influence of a single temporal lag on data patterns.

Transitioning from this foundational perspective, a subsequent, more sophisticated model was crafted. Recognizing the inherent temporal nature of the dataset and the potential clout of historical trends, the modeling approach was expanded to incorporate a richer set of predictors. Specifically, this model drew from multiple lags of the gross amount variable, such as 1-day, 2-day, and notably, the 7-day lags, aiming to encapsulate both immediate past fluctuations and more extended weekly cadences. To further refine the model, it was enriched with the influences of holidays, weekends, and the unique dynamics presented by combined festive and weekend days. Among the predictors, "Day" "Holiday" and "HolidayOrWeekend" were prominently featured, alongside the multi lag variables.

The motivation behind this enhanced approach was to unearth subtler relationships in the data and augment predictive prowess. Leveraging the Bayesian framework, the model offered the advantage of credible intervals for coefficient estimations. Delving into the results, it became evident that specific predictors, such as distinct days of the week and holiday occurrences, bore statistical significance.

Restaurant	RMSE	MAPE (%)
R000	7328.29	15.30
R001	8876.87	22.95
R002	3325.82	8.25
R003	2032.17	8.60
R004	3349.88	9.38
R005	4389.99	14.02

Table 19: Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) for each restaurant in Bayesian models

The table in fig25 presents the prediction performance metrics, for the six restaurants. Notably, Restaurant R003 stands out with the lowest RMSE at 2032.17, indicating precise predictions, and a relatively low MAPE of 8.60%, reflecting a small percentage error. Conversely, R001 exhibits a higher RMSE of 8876.87 and a considerable MAPE of 22.95%, suggesting less accurate predictions and a higher proportional error. These results provide insights into the varying prediction performances across different restaurants, emphasizing the effectiveness of the model in specific cases and potential limitations in others. Looking the estimates shown by the model, the revenue lost during the closing of the lookdown is showed in the table 25.

Restaurant	**Daily Average Loss ()**	**Total Loss ()**
R000	26306.46	1841452.52
R001	45613.76	3192962.87
R002	33586.26	2351038.33
R003	17099.45	1196961.66
R004	32014.58	2241020.90
R005	24412.98	1708908.61

Table 20: Average and total estimated loss for each restaurant

7.10 Summary of models' performances and conclusions for the first period series

In assessing the predictive performance of various models across multiple restaurants, it is crucial to identify the standout performers for each model. A key focus is identifying the top-performing models using the Root Mean Square Error (RMSE) metric. This metric is pivotal in assessing forecasting accuracy, where lower RMSE values are indicative of more precise predictions.

RMSE	**R000**	**R001**	**R002**	**R003**	**R004**	**R005**
ARIMA-Fourier	7501.44	7907.73	3191.25	3376.73	4116.19	**4088.80**
ARIMA-Weekly	7563.87	8528.76	3695.36	3499.84	4160.54	4208.97
Random Forest	7808.21	**7255.13**	6605.05	3392.93	4049.47	5077.39
Prophet	**6853.52**	12694.54	3448.27	3807.94	4116.89	4092.67
Holt-Winters	7120.68	12345.67	3680.11	3958.43	4256.78	4173.26
ETS	7807.77	12157.04	3011.82	3284.72	4115.22	4237.75
TBATS	7127.59	12465.53	**2826.97**	3235.32	4174.68	4139.44
Bayesian	7328.29	8876.87	3325.82	**2032.17**	**3349.88**	4389.99

Table 21: RMSE Values for Different Models and Restaurants

The table28 present the Root Mean Squared Error (RMSE) values for a variety of forecasting models applied to different restaurants, designated R000 to R005. The RMSE, a widely-used metric in time series forecasting, measures prediction accuracy, with lower values indicating more accurate models.

The table reveals interesting insights into model performance. Both the ARIMA-Fourier and ARIMA-Weekly models demonstrate competitive performance across various restaurants. Notably, ARIMA-Fourier achieves the lowest RMSE for restaurant R005, suggesting its potential suitability in certain contexts.

The Random Forest model, recognized for its ability to handle complex patterns, exhibits exceptional performance, particularly in restaurant R001 where it records the lowest RMSE. This highlights its strength in managing intricate data dynamics.

Prophet, a model known for its robust handling of seasonal effects, achieves the lowest RMSE in restaurant R003. This suggests its effectiveness in cases where restaurant-specific patterns are predominant.

On the other hand, the TBATS model, designed to address seasonality and trends in time series data, stands out in restaurant R002 with the lowest RMSE. This underscores its capability in dealing with seasonal variations and complex trends.

Lastly, the Bayesian model shows remarkable adaptability, registering the lowest RMSE in restaurants R004 and R005. This versatility suggests its potential for broad applicability across diverse data sets.

In conclusion, our analysis underscores the fact that there is no universally best model for all scenarios in restaurant sales forecasting. The varied performance of models across different restaurants highlights the critical need for tailoring model selection to the specific data characteristics of each establishment and adopting this customized approach is key to improving the precision and dependability of our forecasts in the ever-changing area of time series analysis.

8 Predictions for the future

In this section, we aim to address the third inquiry pertaining to the forthcoming sales performance of the fourth restaurant. Our objective is to gain insights into the revenue prospects and operational viability. To facilitate the training of our predictive models, we have employed sales data encompassing the period spanning from May 5, 2020, to May 3, 2023. It is noteworthy that this interval encompasses the outbreak of the COVID-19 pandemic; however, there were no instances of complete closure during this timeframe. It is worth mentioning that the geographical area in question experienced varying degrees of pandemic-related restrictions, characterized by color-coded designations of red, yellow, orange, or white. Consequently, our models are equipped to harness the predictive potential of additional categorical variables, such as color-coded indicators or lockdown-related dummy variables.

Furthermore, it is pertinent to note that our dataset encompasses a total of 1034 observations, which proves advantageous for enhancing the predictive accuracy of our models through the utilization of a larger and more diverse dataset differently for the first one, the second series presented 3 missing values. Before the predictive analysis the missing values has been estimated with the r function na.seadec the method involves seasonally decomposing the time series data to handle missing values effectively. Seasonal decomposition often involves breaking down a time series into its underlying components such as trend, seasonality, and residuals.

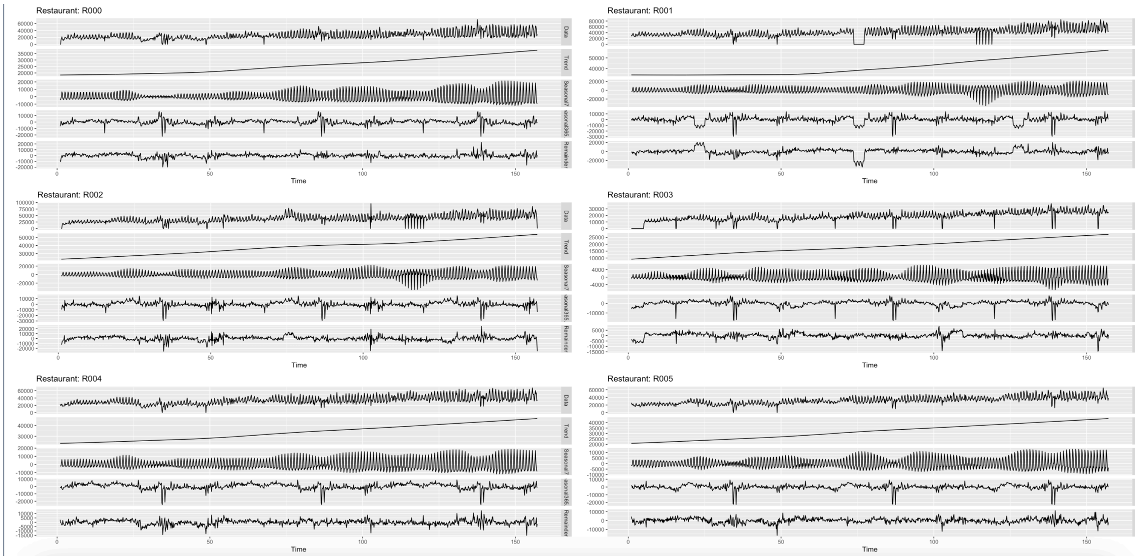

Figure 29: Second series after covid decomposition

We proceeded with the decomposition of the historical series from 4/5/2020 to 3/5/23, considering the data related to sales with daily granularity. This operation was necessary since the historical series they have different patterns inside them. In case examined, the components turn out to be data, trends, daily seasonality and residuals. Within this project the decomposition was performed to improve the understanding of the time series explored by going to study the different patterns that compose it. In this case we have an increasing trend, this indicates that the restaurant is selling more and more over time and with a constant increase throughout the year, so an economic recovery after in Covid-19. The seasonality instead highlighted in the time series during the period between 2020 and 2023 in Italy is strongly influenced by the COVID-19 pandemic. During these three years, the series presents a non-stationary trend, with significant upward and numerous downward peaks. These upward spikes can be attributed to specific events or periods in which demand or economic activity experienced temporary increases. However, it is important to note that the presence of several downward spikes is representative of the negative impact of the pandemic, which has resulted in restrictions and economic uncertainty. This time series clearly reflects the instability of markets during the COVID-19 period, with notable fluctuations in seasonal data. Carefully analyzing these fluctuations can be crucial to understanding the overall effects of the pandemic on economic performance and to adequately plan for the future.

8.1 Arimax Weekly + Red Color Zone

In the realm of ARIMA models, the decision to transition from daily to weekly granularity stemmed from the necessity for more robust modeling techniques. This transition yielded several advantageous outcomes. It became evident that within the dataset, not only did a weekly seasonality exist, as previously identified in the historical series, but there was also a yearly seasonality component. Additionally, numerous irregular data points were attributed to COVID-19-related closures. Consequently, while ARIMA models allow flexibility in altering parameters such as p and q (as outlined in the preceding markdown), accommodating multiple seasonalities posed a challenge in constructing a resilient model. Hence, the strategy of aggregating data on a weekly basis was adopted for several reasons. Firstly, weekly aggregation effectively mitigated data noise. This noise reduction enhanced the models ability to discern underlying patterns and trends, resulting in more accurate forecasts and a reduction in the Root Mean Square Error (RMSE).Also, daily data exhibited rapid and erratic fluctuations. These fluctuations, inherent to daily granularity, could potentially hinder the models capacity to capture genuine trends. The shift to weekly aggregation curtailed the impact of these daily oscillations.Upon careful examination of the data, it became evident that certain datasets displayed a distinct weekly seasonality pattern. Weekly aggregation facilitated the models improved capture of this underlying weekly seasonality, consequently enhancing predictive accuracy.Adopting a weekly data granularity approach contributed to simplifying model complexity. Daily data, due to its intricate daily fluctuations, sometimes led to overly complex models prone to overfitting. Through the use of weekly aggregation, the modeling process was streamlined, mitigating the risk of overfitting. In cases where the inclusion of daily variables proved unfeasible, a novel approach was introduced. A weekly dummy variable, categorized based on color designations, was incorporated. If any day within a given week received a "red" designation due to COVID-19 restrictions, the entire week was categorized as "red." This methodology facilitated the integration of pertinent information while accommodating the shift to weekly data granularity.

Restaurant	RMSE	MAPE (%)
R000	3190.56	7.37
R001	5889.65	10.88
R002	5650.19	10.02
R003	2385.12	6.62
R004	4780.89	9.20
R005	3475.55	8.34

Table 22: Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) for each restaurant using ARIMAX on series post covid

The table 24 summarizes the performance of ARIMAX experiments with 'Color', 'Day', and Lagged Variables for different test sets. Restaurant R003 stands out with the lowest RMSE of 2332.12, indicating precise predictions, and the smallest MAPE of 6.62%, reflecting a minor proportional error. Conversely, R005 exhibits the highest RMSE of 9511.55 and the largest MAPE of 19.34%, suggesting less accurate predictions and a higher proportional error. These results provide insights into the varying prediction performances of ARIMAX models across different experiments, emphasizing their effectiveness in specific cases and potential challenges in others.

8.2 Arimax with Fourier Series

The fundamental principle behind this model is that any intricate cyclic pattern can be expressed as the sum of these basic waveforms, or harmonics. Each sine and cosine function in the Fourier series corresponds to a specific frequency, defined as a harmonic. The motivation behind leveraging multiple harmonics is to accurately depict more intricate seasonality patterns. The first harmonic, represented by the fundamental sine and cosine pair, captures the primary cycle. In the context of a weekly cycle, this primary harmonic can represent the main patterns observed throughout the week. However, real-world data often consists of subtler patternssecondary peaks or troughs within the weekthat the primary harmonic might not encapsulate. To effectively represent these nuanced patterns, additional harmonics have been added. In this specific case, after multiple iterations and evaluations, we found k=3 to be the most optimal. This means that we are considering three harmonics, allowing us to capture and represent the intricate details of our cyclic pattern more precisely.

Restaurant	**RMSE**	**MAPE (%)**
R000	9140.53	20.06
R001	11468.63	15.51
R002	9173.74	14.09
R003	2638.15	7.30
R004	7294.09	13.74
R005	8677.49	17.13

Table 23: Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) for each restaurant using ARIMAX Fourier on series post covid

The table 24 summarizes the performance of ARIMA experiments with Fourier terms, 'Color', and 'Day' dummies for different test sets. R003 stands out with the lowest RMSE of 2638.15, indicating precise predictions, and the smallest MAPE of 7.30%, reflecting a minor proportional error. Conversely, R000 exhibits the highest RMSE of 9140.53 and the largest MAPE of 20.06%, suggesting less accurate predictions and a higher proportional error. These results provide insights into the varying prediction performances of ARIMA models with Fourier terms and additional features, emphasizing their effectiveness in specific cases and potential challenges in others.

8.3 Random Forest

We proceed to train a Random Forest model using the training data with the daily granularity. This model aims to predict the gross amount based on various input features, including "Day," "Weekend," "HolidayOrWeekend", the lag values, and "Color". The model incorporates historical data, categorical features, and lagged values to make predictions. To understand the importance of each feature in the model, we calculate feature importance scores30. These scores help identify which features have the most significant impact on sales prediction.

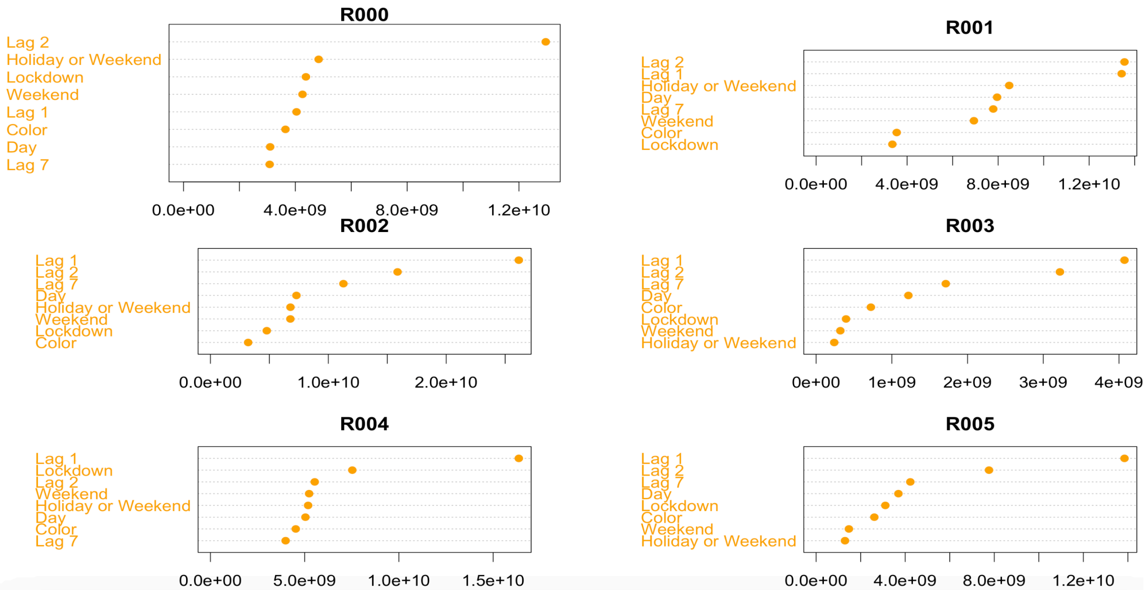

Figure 30: Feature importance in Random Forest Increase in Node Purity as metric

The data highlights significant differences in key variables among restaurants R000 to R005. For instance, the "Day" variable demonstrates a strong influence on R001, suggesting a noteworthy dependency on the day of the week. Similarly, "Lag 2" has a notable impact on R002 and R005, indicating the importance of past performance in determining current success. However, there is variation in the "Weekend" and "Color" variables, with different levels of influence among the restaurants. The presence of "Lockdown" shows variable impacts, with R004 and R000 appearing to be more sensitive to this factor. Overall, the analysis underscores the need to consider the unique characteristics of each restaurant in the context of the random forest model.

Restaurant	RMSE	MAPE (%)
R001	8767.37	17.35
R002	11339.43	15.82
R003	9741.55	15.18
R004	5677.18	19.61
R005	7845.49	14.21
R006	7599.99	14.65

Table 24: RMSE and MAPE for each model using Random Forest on series post covid

The table 24provides an overview of the Random Forest model's performance on different restaurants labeled as R001 to R006. Notably, the RMSE values range from 5677.18 to 11339.43, indicating variations in prediction accuracy across restaurants. Restaurant R004 stands out with the highest RMSE and MAPE of 5677.18 and 19.61%, respectively, suggesting challenges in accurately predicting outcomes for this particular restaurant. Conversely, Restaurant R002 demonstrates a relatively lower RMSE of 11339.43 but a moderate MAPE of 15.82%. Overall, this table offers insights into the model's predictive performance variability across different restaurant scenarios, emphasizing the need for further analysis to understand the factors influencing these outcomes.

8.4 Prophet

We have once again employed the Prophet model in our analysis and enhanced its performance by introducing additional regressors or features. After conducting several trials, we determined that the best combination of features for the model includes "Day", "Color", "Lockdown," and "Weekend." We iteratively incorporated these categorical regressors into the Prophet model using the addregressor function. This augmentation significantly improves the models ability to capture variations in the target variable associated with these categorical features.

Restaurant	**RMSE**	**MAPE (%)**
R001	8690.97	20.49
R002	9391.19	15.82
R003	12106.95	21.33
R004	4136.09	12.51
R005	7539.23	14.53
R006	6223.92	12.55

Table 25: RMSE and MAPE for each model using Prophet on series post-covid

The table 25 presents the prediction results for every restaurant using the Prophet forecasting model. Notably, Restaurant R004 stands out with the lowest RMSE at 4136.09, indicating precise predictions, and a relatively low MAPE of 12.51%, reflecting a small percentage error. Conversely, R003 exhibits the highest RMSE of 12106.95 and a considerable MAPE of 21.33%, suggesting less accurate predictions and a higher proportional error. These results offer insights into the varying performance of the Prophet model across different restaurants in the post-COVID series, emphasizing its effectiveness in specific cases and potential challenges in others.

8.5 Holt-Winters

We initialize a Holt-Winters forecasting model designed to capture both seasonality and underlying patterns in time series data. We specify that the seasonality in our data is "multiplicative," indicating that the seasonal variations are proportional to the overall level of sales. For R004 as example the parameters are:

- **Alpha (level parameter):** $\alpha = 0.082$
- **Beta (trend parameter):** $\beta = \texttt{FALSE}$
- **Gamma (seasonal parameter):** $\gamma = \texttt{FALSE}$

Here, α is the smoothing parameter controlling the weight given to the most recent observations. A lower value of α indicates greater smoothness in the forecasts. We have set β and γ to `FALSE`, indicating that we are not considering a trend component or a damping component in our model.

Restaurant	RMSE	MAPE (%)
R000	11869.57	29.00
R001	13271.59	20.54
R002	12282.23	22.02
R003	4422.06	11.82
R004	10264.82	21.85
R005	9350.06	16.25

Table 26: RMSE and MAPE values for each restaurant using Holt-Winters on the test set.

The table showcases the RMSE and MAPE values for the different restaurants using the Holt-Winters forecasting model on the test set. Notably, Restaurant R003 demonstrates the lowest RMSE at 4422.06, indicating precise predictions, and a relatively low MAPE of 11.82%, reflecting a small percentage error. In contrast, R001 has the highest RMSE of 13271.59 and a moderate MAPE of 20.54%, suggesting less accurate predictions and a moderate proportional error. These results provide insights into the varying prediction performances of the Holt-Winters model across different restaurants, emphasizing its effectiveness in specific cases and potential challenges in others.

8.6 Bayesian linear regression model

In this case, we chose to use only one model because its performance was significantly better than that of a simpler model. Our approach involved systematically exploring different predictor variables, followed by a rigorous selection process based on their respective posterior inclusion probabilities. Ultimately, the variables selected for this model included "Giorno" (day of the week), "Colore" (color), lagged sales values, "Lockdown" status, and "rain" (rainy weather). These variables demonstrated the most favorable outcomes in terms of model performance. Notably, in contrast to the initial time series analysis, we introduced the variable "Color," and interestingly, the variable "Rain" emerged as significantly influential. Furthermore, we conducted a com- prehensive residual analysis, which involved computing residuals to represent the disparities between observed and predicted sales figures. Importantly, our visual inspection of the residuals did not reveal any discernible autocorrelation patterns.

Restaurant	RMSE	MAPE (%)
R000	7956.34	16.21
R001	9333.23	12.67
R002	7615.30	12.20
R003	4980.58	17.37
R004	7647.50	14.65
R005	6096.21	12.63

Table 27: RMSE and MAPE values for each restaurant Bayesian models

The table 27 presents the results using Bayesian forecasting models. Notably, Restaurant R003 demonstrates the lowest RMSE at 4980.58, indicating precise predictions, and a relatively high MAPE of 17.37%, suggesting a moderate proportional error. Conversely, R001 has the highest RMSE of 9333.23, but a lower MAPE of 12.67%, indicating less accurate predictions with a smaller proportional error. These results provide insights into the varying prediction performances of Bayesian models across different restaurants, emphasizing their effectiveness in specific cases and potential challenges in others.

8.7 Results post covid series

Analyzing the results presented in Table 28, it's evident that the ARIMA-Weekly-Red model showcases exceptional performance across all listed restaurants, consistently recording the lowest RMSE values. This remarkable consistency signifies its strong capability in capturing and forecasting sales patterns on a weekly basis.

The ARIMA-Weekly-Red model's dominance is clear as it achieves the best RMSE scores for every restaurant, illustrating its adeptness in handling a variety of sales data across different establishments. Its performance is especially noteworthy considering the complexity and variability inherent in restaurant sales data. The weekly aggregation of data in this model likely plays a crucial role in this success, allowing it to effectively smooth out daily fluctuations and focus on broader weekly trends. This makes it an ideal choice for scenarios where understanding and forecasting these broader trends is more critical than daily sales nuances.

It is also important to acknowledge the commendable performance of the ARIMA-Fourier model. While it does not outperform the ARIMA-Weekly-Red model, it still holds its ground as a strong contender, especially when compared to other models like Random Forest, Prophet, Holt-Winters, ETS, TBATS, and Bayesian. The ARIMA-Fourier model's relatively good performance suggests its usefulness in scenarios where the integration of Fourier terms aids in capturing seasonal patterns that are not strictly weekly.

The Bayesian model, while not the top performer, does show particular strength in forecasting for Restaurants R002 and R005, indicating its potential effectiveness in certain specific contexts.

In conclusion, while the ARIMA-Weekly-Red model stands out as the superior choice for forecasting on a weekly basis, the ARIMA-Fourier and Bayesian models also demonstrate valuable forecasting capabilities in their respective areas. This highlights the importance of selecting a forecasting model that aligns closely with the specific data characteristics and business needs of each restaurant. The choice of model should be guided by the granularity of the data, the specific forecasting requirements, and the operational context in which the forecasts will be applied.

RMSE	**R000**	**R001**	**R002**	**R003**	**R004**	**R005**
ARIMA-Weekly-Red	3190.87	5889.76	5650.36	2385.84	4780.54	3475.97
ARIMA-Fourier	9140.44	11468.73	9173.25	2638.73	7294.19	8677.80
Random Forest	8767.21	11339.13	9741.05	5677.93	7845.47	7599.39
Prophet	8690.52	9391.54	12106.27	4136.94	7539.89	6223.67
Holt-Winters	11869.68	13271.67	12282.11	4422.43	10264.78	9350.26
ETS	8439.77	12684.04	12215.82	4883.72	6729.22	8332.75
TBATS	10004.59	13336.53	13442.97	4370.32	8009.68	9384.44
Bayesian	7956.29	9333.87	7615.82	4980.17	7647.88	6096.99

Table 28: RMSE Values for Different Models and Restaurants

9 Rolling forecasting origin

Following the guidelines outlined in our methodology (see Section 6.1), we applied a rolling origin forecast analysis for the periods of 1, 7, 14, 30, and 60 days. In this study, we focused exclusively on ETS model, as it demonstrated competent generalization across all time-series under investigation. While the ETS model was not the top performer in every individual case, its consistent performance across different datasets, coupled with its simplicity and interpretability, made it a preferable choice for our analysis. This decision to employ the ETS model was grounded in our objective to balance predictive accuracy with model simplicity, facilitating easier interpretation and implementation in practical settings.

Restaurant	RMSE 1 Day	RMSE 7 Days	RMSE 14 Days	RMSE 30 Days	RMSE 60 Days
R000	9093.388	10656.848	11071.585	12820.940	13224.619
R001	11268.07	13021.58	13158.90	13744.16	16322.90
R002	11173.62	12467.26	12795.03	13518.04	13717.64
R003	2574.614	3133.827	3177.497	3388.061	3972.694
R004	9358.917	10458.447	10461.374	11027.915	11189.034
R005	6462.021	7475.798	7668.955	8023.361	9173.506

Table 29: RMSE for every restaurant in horizon 1, 7, 14, 30, 60

Restaurant	MAPE 1 Day	MAPE 7 Days	MAPE 14 Days	MAPE 30 Days	MAPE 60 Days
R000	31.39951	30.34687	31.56890	41.23862	41.62308
R001	23.33059	24.15456	24.27048	24.89688	29.43380
R002	24.65807	25.16161	26.17954	28.01331	27.68193
R003	9.805369	9.407571	9.615743	11.071543	11.957314
R004	23.54393	22.90451	22.85825	23.67147	24.11263
R005	15.98031	16.08054	15.65287	16.22952	20.73742

Table 30: MAPE for every restaurant in horizon 1, 7, 14, 30, 60

The RMSE table **??** offers insights into the model's performance over forecast horizons of 1, 7, 14, 30, and 60 days. Notably, the RMSE values increase with the length of the forecast horizon, which is a common occurrence in time-series forecasting due to the accumulation of errors over time. For instance, Restaurant R003 consistently shows the lowest RMSE values across all horizons, indicating a higher degree of accuracy in forecasts for this particular establishment. In contrast, Restaurants R000 and R001 exhibit higher RMSE values, especially in the longer horizons (30 and 60 days), suggesting more significant challenges in accurately forecasting over extended periods. The MAPE table**??** complements this analysis by providing a percentage-based error metric, which offers a more intuitive understanding of the model's performance relative to the scale of the data. For instance, the lower MAPE values for restaurant R003 across all horizons reaffirm its superior forecasting accuracy. Conversely, restaurant R000 demonstrates higher MAPE values, particularly in the 30 and 60-day horizons, underscoring the increasing difficulty of making accurate long-term forecasts.

10 Conclusions and future developments

Following the analyses conducted, several conclusions can be drawn regarding the research questions posed.

The initial exploratory analysis was instrumental in identifying the unique characteristics of each time series data available. This was crucial for the development of predictive models. Additionally, a significant and precise data enrichment process was carried out, tailored to the specific location of each restaurant. This included the addition of many important attributes, such as the zone color, lockdown status, and holiday periods. This phase of the analysis not only provided valuable data for predictive modeling but also facilitated a comprehensive exploratory study, evaluating the impact of climatic conditions and other factors on revenues.

A key goal of this research was to estimate the financial losses experienced during the early phase of the pandemic. To this end, various models were evaluated and the most effective and reliable ones were chosen based on the data at hand. The forecasts generated by these models confirmed the substantial impact of the 2020 lockdowns, highlighting missed revenues both for the entire period and on a daily basis across all restaurants. It was observed that while some restaurants quickly recovered, others did not.

Furthermore, an estimation of sales for the two months following the last data point was undertaken, considering the unique characteristics of each restaurant. Finding a model capable of accurately predicting revenue trends is a valuable tool for market analysis, facilitating better resource management. It was found that there is no universal model that excels for all restaurants; rather, different models are suited to different establishments, depending on their specific characteristics. Each restaurant's data was meticulously analyzed and interpreted with its unique context in mind.

This research can be expanded in several ways. For example, if price lists were available, it would be possible to determine which products experienced the greatest price increase, and whether the post covid rise in receipt prices was due to specific products or a general pricing policy across the restaurants. The accuracy of the forecasts could be improved with access to a larger quantity of historical data and additional features. This would also enable the use of more sophisticated deep learning models, such as transformers, which require complex and extensive datasets. With high-quality, restaurant-specific data, more precise, tailored analyses could be conducted. These might include examining staff performance, turnover, and effectiveness, as these factors can directly influence sales. Additionally, analyzing customer feedback during specific periods could provide insights into a restaurant's performance. Further, considering data on food trends and changing consumption habits could offer valuable perspectives.However, in this study, the identities of the restaurants remain anonymous; only their locations are known, so features were added based on available information. With a larger set of restaurants, clustering analysis could be performed to identify groups of restaurants with similar characteristics. This approach could uncover hidden patterns in the data, enhancing understanding of the differences and similarities among various establishments.

This study establishes a basis for more detailed future investigations and the development of more focused business strategies, underscoring the significance of data analysis in the restaurant industry.

Bibliography

[1] Abdulkader Aljandali. The box-jenkins methodology. pages 59–79, 2017.

[2] Leo Breiman. Random forest, vol. 45. *Mach Learn*, 1, 2001.

[3] Patrick Burns. Robustness of the ljung-box test and its rank equivalent. *Available at SSRN 443560*, 2002.

[4] Alysha M De Livera, Rob J Hyndman, and Ralph D Snyder. Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American statistical association*, 106(496):1513–1527, 2011.

[5] Rob J Hyndman and George Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2018.

[6] Prajakta S Kalekar et al. Time series forecasting using holt-winters exponential smoothing. *Kanwal Rekhi school of information Technology*, 4329008(13):1–13, 2004.

[7] Mathieu Lepot, Jean-Baptiste Aubin, and François HLR Clemens. Interpolation in time series: An introductive overview of existing methods, their performance criteria and uncertainty assessment. *Water*, 9(10):796, 2017.

[8] Ricardo P Masini, Marcelo C Medeiros, and Eduardo F Mendes. Machine learning advances for time series forecasting. *Journal of economic surveys*, 37(1):76–111, 2023.

[9] Meteo.it. Meteo.it - previsioni meteo, 2023.

[10] Aakash Parmar, Rakesh Katariya, and Vatsal Patel. A review on random forest: An ensemble classifier. In *International conference on intelligent data communication technologies and internet of things (ICICI) 2018*, pages 758–763. Springer, 2019.

[11] Ďurka Peter and Pastoreková Silvia. Arima vs. arimax–which approach is better to analyze and forecast macroeconomic time series. In *Proceedings of 30th international conference mathematical methods in economics*, volume 2, pages 136–140, 2012.

[12] Robert H Shumway, David S Stoffer, Robert H Shumway, and David S Stoffer. Arima models. *Time series analysis and its applications: with R examples*, pages 75–163, 2017.

[13] Ivan Svetunkov and Fotios Petropoulos. Old dog, new tricks: a modelling view of simple moving averages. *International Journal of Production Research*, 56(18):6034–6047, 2018.

[14] Leonard J Tashman. Out-of-sample tests of forecasting accuracy: an analysis and review. *International journal of forecasting*, 16(4):437–450, 2000.

[15] Sean J Taylor and Benjamin Letham. Forecasting at scale. *Facebook, Menlo Park, California, United States*, 2017.

[16] Johannes Tshepiso Tsoku, Nonofo Phukuntsi, and Daniel Metsileng. Gold sales forecasting: The box-jenkins methodology. *Risk Governance and Control: Financial Markets Institutions*, 7:54–60, 2017.

[17] Stylianos I Vagropoulos, GI Chouliaras, Evaggelos G Kardakos, Christos K Simoglou, and Anastasios G Bakirtzis. Comparison of sarimax, sarima, modified sarima and ann-based models for short-term pv generation forecasting. In *2016 IEEE international energy conference (ENERGYCON)*, pages 1–6. IEEE, 2016.