MIE1623- Assignment 2 Forecasting

Team members:
Hajra Begum Student number: 1005614015
Meghana Padmanabhan Student number: 1000905210
Jiayue Niu Student number: 1001122894
Amir Kharazmi Student number: 1000909332

Excel workbook name: StJuliusHospital_CSection_FinalVer.xlsx

1. What data cleaning did you do?

Handling Missing/Bad Data

Generally in terms of data cleaning, for missing or bad data, there are two approaches possible: Deletion or Imputation (i.e. replacing the missing data with some replacement value). There are 4 columns with missing data:

- **CS Type:** 52 missing records (2.4% of the total)
    - Given the relatively small percentage of missing records (<5%) and the random nature of these records, we remove these rows.
- **Total OR Time/OR Departure Time/OR Arrival Time:** 205 records (9.3% of total).
    - OR Departure/Arrival Time are not data points which we'll use directly (they are mainly used to get total OR time or day/month/week of the year which we can get separately from Delivery Date and Delivery Time columns). As for OR Arrival time, we'll use the average OR time to replace missing/NULL values.

Creating New Columns

In addition to handling missing values, we also create new features/columns to explore seasonality and themes within the data. Below are a few columns that we have added as well as the reasoning for each addition:

- **Weekday vs. Weekend**: There will likely be much fewer deliveries on weekends vs. Weekdays (the OR is only doing unplanned operations on weekends). We wanted to see if there is a material difference between the number of unplanned operations on weekends vs. weeks.
- **Day vs. Night**: We classified each day into Day (6AM to 6PM) or Night (6PM to 6AM) categories. OR rooms run between 9-5PM and hence all night deliveries are unplanned. We'd like to see if there are any spikes in terms of night deliveries on any specific day/month.
- **Week of the year:** We expect a slight uptrend (i.e. population is growing) in week-over-week total number of births. Additionally, when looking at total number of

births, in order to deal with the volatility of daily data, we look at weekly births instead to smoothen the analysis.

- **Month of the year:** We do not expect a material difference between each month's data, however we have also added a column for month of the year to allow us to aggregate the data over each month.

2. Does the cleaned data display seasonality or trends? And 3. What forecasting methods are appropriate for this type of data?

Step1:

Data Type:

According to the data given it is the set of observations over time.So,Time series Analysis is recommended.

Step 2:

Look For Pattern:

- Stationary (flat)
- Linear trend
- Seasonal
- Seasonal with trend

Inorder to look for pattern we needed to identify data points to be plotted Vs Time.

We collected the most important data based on the interpretation from the problem description. (i.e hospital struggling to reach their target of performing 98% of scheduled C-sections on time due to the unpredictability of emergency C-sections).

After performing data Cleaning we considered Planned and Unplanned surgeries to be the most important categories for our analysis.These are our data points(Y values) in Table1 to be plotted/considered.Out of 2210 data points we are considering the 2158 data points for analysis with complete information.

Count of Patient	Column Labels	
Row Labels	Planned	Unplanned
Planned-as scheduled	897	
Planned-not as sched	213	
Unplanned-		2
Unplanned-crash		74
Unplanned-urgent		972
Grand Total	**1110**	**1048**

Table1:Planned Vs Unplanned

Next ,we needed a range of X values(Time).To find out which is the suitable metric for X from the options(Days,Weeks,month,seasons) .

Plot1:

We used a line plot to check the pattern of (planned vs unplanned) data along Days(of the week).

In Figure 1 we didn't see any trend and seasonality ,so to make sure other metrics of time show us better patterns we didn't consider this plot for further analysis.We also observed that

predicting Unplanned data was more important for the analysis as it corresponds to the emergency births and has a inverse effect on the Target of achieving more planned births(scheduled).

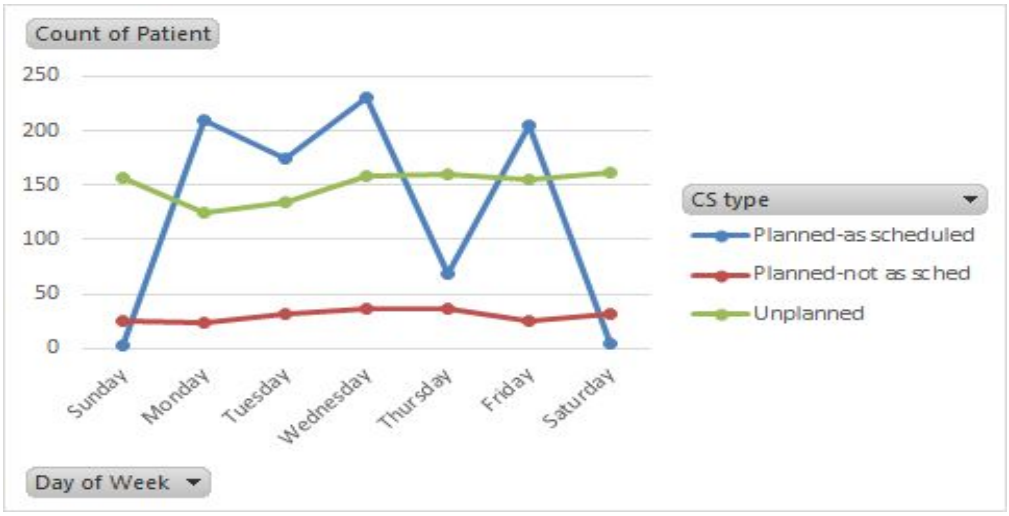

Figure1:Count of patients(planned and Unplanned) by days of the week

Plot 2:

We then plotted unplanned data (Y- values) by weekdays for months and did not consider Weekends as unplanned surgeries during weekdays affect the planned births.

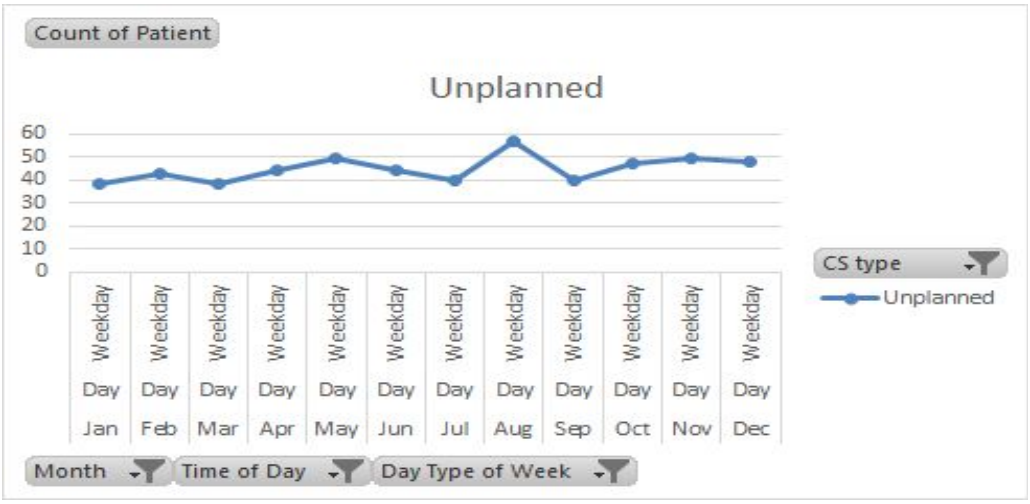

Figure2:Count of patients(Unplanned) by weekday for month

In figure 2 the plot didn't account for any seasonality and trend and was stationary.We then plotted again to check if we could find any other patterns in the data.

Plot 3:

Our team then decided to plot weekdays by days of the week for months for unplanned C-sections data points.

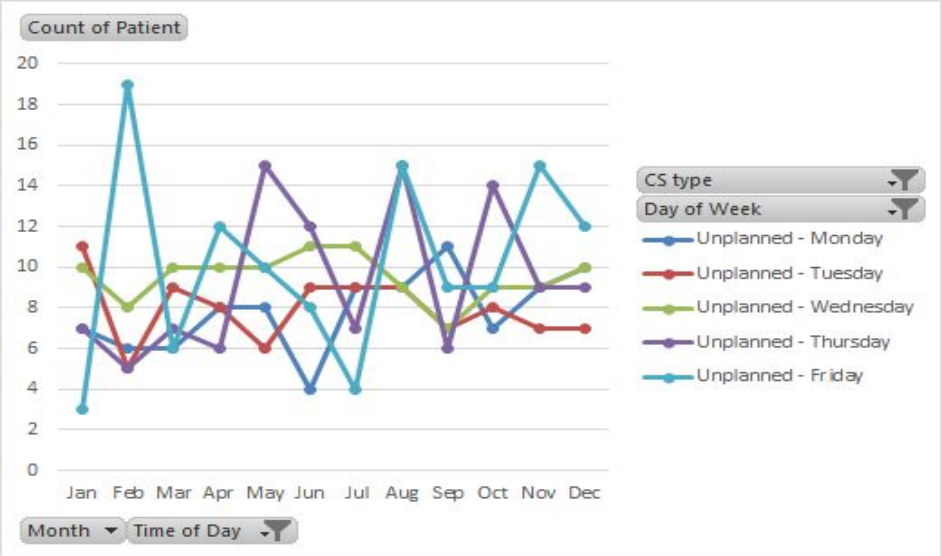

Figure3:Count of patients(Unplanned C-section s) by days of the week for month

In Figure 3 we found that the the trend was stationary and there was no seasonality.

Plot 4:

Finally we plotted the graph for unplanned (emergency) births that happened on the weekday (Monday- Friday) for weeks of the year(1-52).

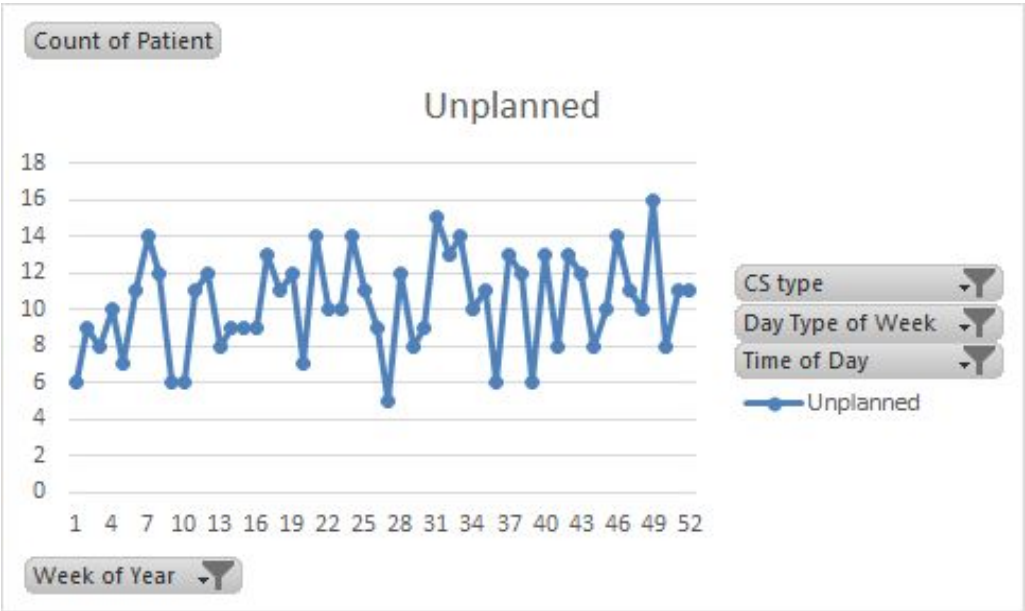

Figure4:Count of patients(Unplanned) on the weekdays for weeks of the year

In this figure we observe that the data is stationary in terms of trend and there is no seasonality pattern.
After looking into different graphs we decided to do forecasting using weekly data because monthly data is too general and also hospitals rarely do forecast everyday. Predicting demand on a week basis is more reasonable and applicable.

Step 3:

Forecasting methods

	Stationary	Linear trend	Seasonal	Seasonal + trend
Linear regression	✔	✔		
Moving average	✔			
Weighted MA	✔			
Exponential smoothing	✔			
Double MA		✔		
Holt (double ES)		✔		
Seasonal exponential			✔	
Holt-Winters (triple ES)				✔

As our time series is stationary and no seasonality we decided to use Stationary-Forecasting methods and compare them to get the best results.
We also considered that our data was only for one year(i.e 52 weeks) there may be some increasing or decreasing linear trend that couldn't be obtained from plotting limited data.So our team decided to use some of the Linear forecasting methods also to compare and to predict any hidden linear trend(if better result could be obtained from linear-Forecasting methods).

	Stationary	Linear
Linear Regression	√	
Moving averages	√	
weighted moving Averages	√	
Exponential smoothing	√	
Double moving Average		√
Double Exponential smoothing		√

Table2:Implemented forecasting methods

4. What is the MAD and MSE of each of your forecasting methods?

Forcasting Methods	MAD	MSE
Moving Average (n=3)	2.47	9.32
Moving Average (n=6)	0.76	7.86
Moving Average (n=9)	0.41	6.14
Weighted Moving Average(n=3)	2.47	9.28
Weighted Moving Average(n=6)	2.11	7.72
Weighted Moving Average(n=9)	1.66	4.98
Exponential Smoothing	2.34	8.37
Double Moving Average(n=1)	3.78	24.16
Double Moving Average(n=3)	5.22	40.68
Double Moving Average(n=6)	7.21	77.76
Double Moving Average(n=9)	9.63	129.56
Holt (Double ES)	2.47	8.74
Linear Regression	2.14	6.59

Table 3: Results of Forecasting Methods

Notes: Parameters used were either optimized by solver or tuned to get good performance. Exact parameter values used in forecasting can be found in the corresponding spreadsheet.

Conclusion: Moving Average (n=9) has the best MAD while Weighted Moving Average(n=9) has the best MSE.The result wasn't surprising because our demand(number of unplanned surgeries) is quite stationary; the forecast gets better by considering more past data.

5. Do you trust your forecasts? Why or why not?

Yes, but only to some extent. We trust our forecasts because MAD and MSE are low using Moving Average and Weighted Moving Average. These methods can give useful predictions that are close to actual values. For example, with 52 weeks of data, the weekly predictions were only off by 0.41 patient count on average. However, the forecasts cannot always be trustworthy because there may be changes/advances in technology, policy, and other sectors. This may add trend or seasonality to the demand data distribution. In that case, we should change our forecasting approach. Lastly, there is always uncertainty and error that cannot be eliminated from forecasts by any means. We should keep this in mind especially when dealing with large scaled data. To conclude, our forecasting results may be wrong or off by a little but they provide support in making decisions.

6. When do you recommend that SJH build a fourth OR, if at all?

Figure 5: Rate of of scheduled C-sections on time by month

It is evident in Figure 5 that on-time rates of scheduled C-sections fail to meet the 98% target. The major cause for the postponement or cancellation of scheduled C-sections is random emergency C-sections. Thus, adding a fourth OR room can help cover this unexpected demand and further prevent planned surgeries from getting delayed or cancelled.

However, referring to Figure 4, there are around 10 unplanned C-sections during regular hours (9am-5pm) on weekdays(Monday-Friday) every week. By applying filters to the cleaned data in the excel sheet, there are 1806 surgeries done in 52 weeks during regular hours on weekdays with the current 3 OR rooms. Therefore, each OR room can cover up approximately 1806/(52*3) ≅12 C-sections on average regardless of planned or unplanned. Theoretically and ideally, this will be able to resolve all the unplanned cases during day time on weekdays. However, SJH should first evaluate how much benefit it will bring or how much cost it can avoid to have more scheduled C-sections on time and compare with benefit brought by other services that could have been improved using the money. If increasing on-time rates for C-sections is the priority, SJH should start building the fourth OR room as soon as possible. Otherwise, it may be better to spend on other services since building another OR room just for C-sections based on the forecasted demand which is very less, may result in a low utilization rate of that fourth OR room.