Jakarta's Ambulance Service Call Volume Prediction: A Forecasting Approach

Pavan Akula Venkata Gnyana
Student#: 23078907

1. INTRODUCTION

Based on previous data patterns of **call volume per day**, the project's goal is to forecast the number of calls to Jakarta Ambulance Service (118) for two months, from October 2019 to November 2019. The prior data was made available between March and September of 2019. We look for trends, seasonality, and other patterns using time series analysis and basic to advanced forecasting models and methodologies in order to accurately forecast call volume per day with this analysis.

A variety of time-series forecasting models are examined in this poster. Selecting the best model to represent and forecast the call volume data requires starting with a baseline Naïve model, a simple mean model, a moving average model, moving through exploration models including SES, Holt Linear, Holt Winters, and simple linear regression, and concluding with a suite of ARIMA.

2. DATA PRE-PROCESSING

Column	Non-Null	Count	Dtype
Call_Date	32709	non-null	datetime64[ns]
Call_Volume	32709	non-null	int64
Date	32709	non-null	object
Time	32709	non-null	object
Day	32709	non-null	int32
Day_Name	32709	non-null	object
Week	32709	non-null	UInt32
Weekday_Weekend	32709	non-null	int32
Month	32709	non-null	object

Firstly, raw data is loaded into Python, and **196 duplicates were removed** to maintain the uniqueness as the date-time was matching exactly, including seconds, which is unusual. The data pre-processing is done by the Pandas package, where initially date-time is split and Day, Day Name, Week, Weekday/ Weekend, and Month are extracted for better visualisation as per the above table. Also, a small check was performed to see if there were any missing dates between the date range so that there are no gaps in the data while plotting and to draw the inference that the data is continuous.

3. DATA SCRUTINIZATION

Let's see the numerical and graphical summaries of the data for better understanding before feeding it to different models.

3.2 Numerical Summary

Index	Call_Volume
count	214
mean	152.845794
std	36.994094
min	60
25%	119
50%	163
75%	183
max	210
sum	32709
var	1368.562963
skew	-0.402678
kurtosis	-1.037834
co-var	24.15%
Shapiro–Wilk test(P-Value)	7.55296E-08

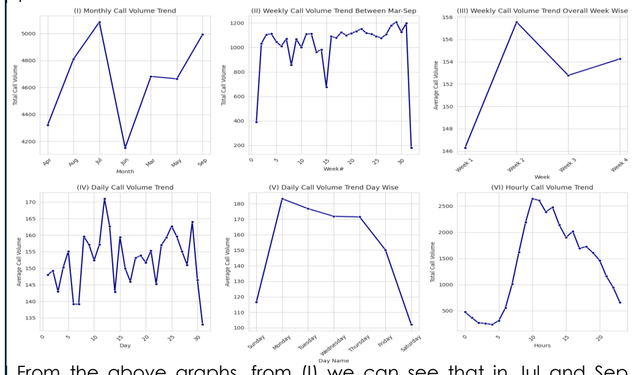
The above table is the numerical summary of calls made to the ambulance service per day. We can see that, on average, around **153 calls** were made **per day**. By referring to the P-value of the Shapiro-Wilk test as less than significance level 0.05, we can say that the **data is not normally distributed**.

The skewness is -0.402, which suggests that the data is slightly skewed to the left. Also, the kurtosis is approximately -1.04, which further indicates a platykurtic distribution, i.e., lighter tails than a normal distribution. The coefficient of variation (co-var) is approximately 24.15%, indicating the relative variability of the data compared to its mean.

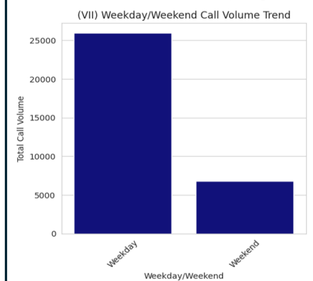
As we saw the basic numerical summary of the calls made, now let's see the graphical summary where the call volume is drilled down from month level to time level with the help of the pre-processed data and further visualise different versions of moving averages which helps us to choose the window size for smoothening of the data.

3.2 Graphical Summary

Now, let's see the graphs that depicts the call volume trend over period of time that further visualises the behaviour of data.



From the above graphs, from (I) we can see that in Jul and Sep there were more calls and in Jun there were less when compared to others. From (III) and (V), a greater number of calls were made in Week 2 and on Monday respectively each month. Finally, from (VI) the call volume was very high between 8:00 to 22:00.



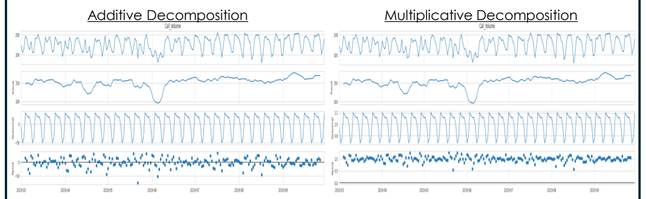
The Bar Plot shows the comparison of call volume between weekdays and weekend, and we can see that a greater number of calls were made on weekdays by which we can conclude that there is higher call activity in weekdays during the week with fluctuations observed on the above graphs on specific days.

4. DECOMPOSITION AND MOVING AVERAGES

The data decomposition and moving averages are some of the essential techniques in time series analysis for understanding and modelling the underlying patterns and trends in the data.

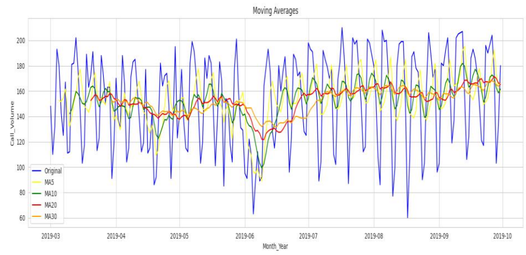
4.1 Additive and Multiplicative Decomposition

The below plots represent the decomposition of the data into different components that are separated and visualised.



Here, the observed time series is decomposed into three components, where the first row represents the original data and the rest three are trend seasonality and residuals, respectively. **The trend component** of additive and multiplicative decomposition represents the long-term direction of the data, whether it is an upward trend or a downward trend, but in our case, there is a minimal trend. **The seasonal component** constitutes the seasonality of the data, where there are any repeated patterns seasonally in the data. Here, we have weekly seasonality among the months over the entire period. **The residual component** provides valuable insights to the unexplained variations and random fluctuations present in the time series data while forecasting the future values. The data shows good model fit, with evenly scattered residuals above and below the trend line, capturing systematic patterns and leaving random fluctuations, ensuring reliability for analysis and forecasting.

4.2 Moving Averages



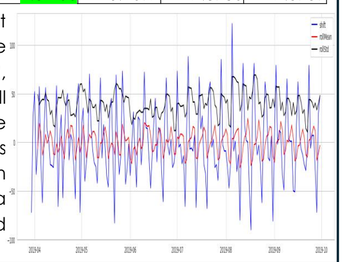
Moving averages are commonly used to smooth the data fluctuations and highlight the underlying trends or patterns in time series analysis. In this project, moving averages of 5, 10, 20, and 30 are used to smooth the short (5), medium (10), long (20), and longer (30) fluctuations and to view short-term, medium-term, long-term, and longer-term trends or patterns. The longer the moving average window, the greater the stability and long-term view of the underlying trend. But in our case, the data is limited to 7 months; hence, a maximum of 30 days window is considered.

5. DATA STATIONARITY

The term stationarity states that the data is stationary if there is no trend or seasonality in the data, i.e., there is a constant mean, constant variance, and constant autocorrelation with respect to time. It is always a good practice to make the data stationary and then feed the same to the time series models. There are several techniques to make the data stationary. Here, we chose to do first-order differencing as it is one of the most effective ways to make the seasonality data stationary and simple to unshift once the model is trained. To check the stationarity of data, there are many statistical unit root tests, and among them is the Augmented Dickey-Fuller Test. Below are the test results:

Augmented Dickey–Fuller Test					
	Test Statistics	P-Value	1%	5%	10%
Before Time Shift	-1.994988247	0.28877	-3.4694	-2.8787	-2.5759
After Time Shift	-5.526636132	1.8E-06	-3.4692	-2.8786	-2.5759

In the results, we can see that the P-value is less than the significance level (alpha = 0.05), so we can reject the null hypothesis, which says that the data is not stationary. Now, as the data is stationary, we can proceed with feeding this data into the models and find a good fit for forecasting future values.



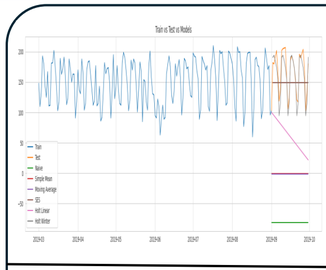
6. MODEL SELECTION

As all the prerequisites are completed including stationarity check, let's proceed for model selection to forecast 2 months call volume.

6.1 Baseline, Simple and Complex Models

The baseline models chosen for this project is with Naïve forecast model and simple approaches like simple mean model, simple moving average model and simple linear regression. Few of the complex models (Extrapolation models) like SES model, Holt Linear model, Holt-Winters model was also chosen but as we move on ahead starting from the baseline model, the Mean Squared Error (MSE) is gradually decreasing which is a good sign for us to forecast future values. The resultant table contains all the error statistics like MSE, MAPE and MAE based on which we can say whether a particular model is healthy for the prediction or not.

Model	MSE	MAPE	MAE	AIC
Naïve Forecast	8385.333	628.97	83.8	N/A
Simple Mean	1372.163	17.14	27.432609	N/A
Moving Average	1383.444	40.28	27.631111	N/A
SES Model	1502.339	21.33	35.704938	1339.089
Linear Regression	1141.901	NaN	27.495332	N/A
Holt Linear	12774.47	61.47	105.376885	1383.879
Holt Winters	245.543	7.84	12.635824	1148.714

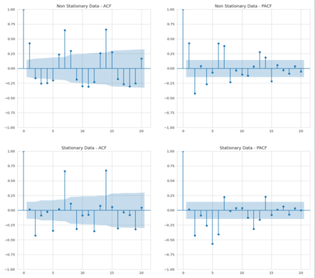


The whole dataset was split into train and test where 6 months data is included in training set and 1 month data is included in test set hence the same is used in all the models. Among these models we can see that Holt Winters model has the least MSE value but further one of the ARIMA model was chosen.

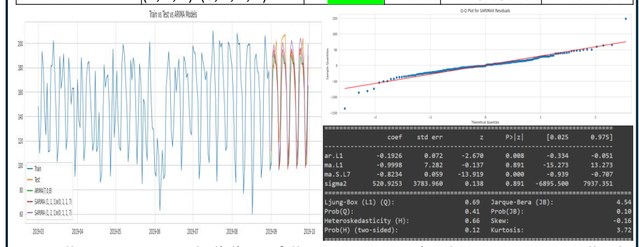
6.2 ARIMA (ACF & PACF)

ARIMA is one of the popular models that is used even today as it is flexible and can be adapted to complex data by forecasting efficiently. But the prerequisite to use ARIMA model is the p, d, q value so to find out these values we must plot Auto Correlation Function (ACF) by which we get the q value and Partial Auto Correlation Function (PACF) by which we get the p value and finally d value is the difference used while making the data stationary.

The ACF and PACF plots are plotted for both non-stationary data as well as stationary data, and while choosing the best ARIMA variant, we can either start with 2 till 9 (2,5,7,9) as a q value and start with 2 till 7 (2,4,5,6,7) as a p value, or an auto_arima function can be used to skip the manual process of choosing p and q. When the auto_arima function was used, the values p = 0 and q = 1 were obtained; hence, we started with the same.



Model	Order	MSE	MAPE	MAC	AIC
ARIMA	(7, 0, 8)	314.432	9.39	15.097538	1676.619
SARIMAX	(1, 1, 1)x(0, 1, 1, 7)	192.645	7.33	11.57581	1586.805
SARIMAX	(1, 2, 1)x(0, 1, 1, 7)	152.824	6.63	10.603198	1624.888

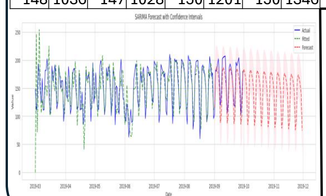


From the summary statistics of the ARIMA variants, we can see that the SARIMA model with the order (1,1,1) and seasonal order (0,1,1,7) has the second lowest MSE. The SARIMA model with the order (1,2,1) and seasonal order (0,1,1,7) has the least MSE, and the residuals are linear. Hence, it is used for forecasting two months of future data.

7. RESULTS AND CONCLUSION

Finally, the forecast for October and November of 2019 is below where there is average and sum of the forecast month-wise and week-wise.

Oct-19								Nov-19							
AVG - 149 ; SUM - 4611								AVG - 139 ; SUM - 4151							
Week 1		Week 2		Week 3		Week 4		Week 1		Week 2		Week 3		Week 4	
AVG	SUM	AVG	SUM	AVG	SUM	AVG	SUM	AVG	SUM	AVG	SUM	AVG	SUM	AVG	SUM
148	1036	147	1028	150	1201	150	1346	143	1000	142	989	140	1120	131	1042



The project successfully explored datasets, built models, and evaluated performance, with SARIMA emerging as the most accurate model. Further refinement and exploration of advanced forecasting techniques could enhance predictive accuracy.