Abstract:

This report includes the results of all of the observations and forecasting I have done as a means to answer the requirements set out in the project proforma. In this report I discuss the methods of observation of the data, Naïve modelling, and also the method of forecasting I chose, ARIMA modelling, and why I chose these forms of modelling to achieve the required goal. It also includes both the advantages and disadvantages of using these models and why I actively chose to use these over other models such as Linear Regression.

Introduction:

This report will go over what methods and observations I used in order to achieve the requirements given by the manager of the local walk in centre, further into the report I will discuss how I did this but the major ways is by using basic statistical concept modelling to understand trends and to get a vague forecast, then using ARIMA as the primary forecasting method which was successful in creating a forecast of the given data.

Statistical modelling:

As a means to begin the process of predictions and forecasting I decided to begin with basic statistical concept modelling, I did this as it will allow me to gain an understanding of both the data itself and any basic patterns that show from it. It also allows me to give the manager a deeper understanding of the data and how some prediction were created as the manager should be able to understand basic statistical concepts.

Using this I found that the busiest periods are actually at the start and ending of each year, specifically the first 2 months and then the 2 last month of each year show the highest amount of people. This is seen throughout each year but is more pronounced in the fourth year:

Chart

Description automatically generated

Figure 1, naïve forecast of number of people in the fourth year

Here we can see the pattern that is in all of the completed years with people increasing significantly at both sides. This means that in order to plan the staff rota more effectively the manager should put more of his workforce at the first 3 months and the last 2 months as to allow for the period where it both ramps up and down. We also see an increased trend of number of patients as the years progress with the fifth year showing the highest it has ever been.

In order to decide which form of forecasting to use I compared each MAPE score of SES, HOLT and ARIMA and ARIMA had the lowest MAPE and therefore was chosen to be used for predictions. I also thought about using Linear regression but after observing the data I decided not to as there would not be a statistically reasonable result since there is so much variance in the data.

In order to use ARIMA we have to remove seasonality and trends, we do this by using log. Before doing this we see a test-statistic of 1.7666 and after this we see a test-statistic of 0.0426 which is a significantly better test-statistic and reliability as it is less than 1. Using the output of the previous function we also find that the lag of the data is 7.

Chart, line chart

Description automatically generated

Figure 2, Graph depicting seasonality removal using single and double differenced logs

Here we see the ACF and PACF before the forecast and as you can see the large majority of lines are within the threshold with only a couple standing out. This will mean that there will be some inconsistency with the data but not enough to massively change the forecast, therefore I believe this data is a good enough fit for ARIMA.

Diagram

Description automatically generated

Figure 3, Graphs displaying the residuals, ACF and PACF of the stationary data

Below is the ARIMA forecast with the order (1,1,7), this gave the most accurate output and therefore will be kept for the forecast. After using the ARIMA with Seasonal turned to FALSE I then decided to see if it would have a better accuracy with Seasonal turned to TRUE but had the same result and therefore it was not required.

Chart

Description automatically generated

Figure 4, Graph showing the forecasted outcome of the ARIMA modelling

From this graph above we can easily see the predicted 7 day period ahead, as you can clearly see the forecast shows that we should see a decrease in the amount of people coming into the local medical walk in with the hard prediction showing only a little drop in people and the softer prediction showing a more significant drop.

The 7-day forecast using the ARIMA model produces the following number of patients: 47, 36, 50, 31, 41, 25, 35 for each day respectively.

Conclusion:

In conclusion I have successfully created a forecast of the 7 days which fits the original data and also fits the trends as previously shown in the data. I have also succeeded in explaining how the rota of the employees should be changed in order to have the most effective output. There are however some advantages and disadvantages that come with using models such as ARIMA and simple statistical modelling such as naïve, these include the following; The Naïve model was very easy to implement and was a very good way to introduce myself to the data and any trends that follow, it is however very poor at considering any possible underlying relationships in the forecast variable. I believe that the implementation of the naïve model mitigates this disadvantage since I was only using it as a means to give a clear understanding of the data and therefore no underlying relationships would have been required

The advantages of the ARIMA are significantly more prevalent here as the dataset provided and the requirements allow it to, this is shown since one of the biggest strengths of the ARIMA model is that it performs well on short term forecasts and since the required forecast was only a week it meant that accuracy was significantly increased. Another advantage is that it only requires previous time series data to forecast which is the only data we actually have. The ARIMA model does have its downsides however and this is shown as it does cannot be used on seasonal time series data and therefore I have had to remove seasonality from the dataset, this may however be a good thing as the model can not use false relationships that may not be there in order to predict. Another disadvantage is that it is computationally expensive and although in this project this was not an issue it may become one if more computational power was required.