Data Analysis using Time Series on overseas trips, new house registration and using binary logistic regression on childbirth.

Priyanka

School of Computing
National College of Ireland
Dublin, Ireland
x20192037@student.ncirl.ie

Abstract— The paper aims to perform binary logistic regression and time series analysis on three datasets from the Central Statistics Office Ireland. The objective of analyzing the Childbirths(dataset1) a) Predict the impact of mother on child's weight during birth. The resultant the binary logistic regression model has 97.3% accuracy that predicts the impact of mother age, a smoking habit with number of cigarettes in a day, Gestation(weeks) of the child on its birth weight. This paper focuses on the health care sector for improving the standard of treatment given to pregnant women in Ireland. Also, Ireland is a developed country where people from all over the world plan there trips as the tourism market are growing rapidly over the years in Ireland and contributing to the nation's revenue. The paper aims to review quarterly overseas trips (dataset 2) b) to predict the impact of season on tourists in Ireland and the last dataset new house registration in Ireland, b) predict the impact of time on a new house registration (dataset 3) in Ireland. The Box-Jenkins models produce accuracy of the prediction and visualizing the trend and seasonality using the Mean forecast, ARIMA, Naïve, Seasonal Naïve, Simple exponential smoothing metrics to validate the accuracy of the performance model.

Keywords— Box-Jenkins, Auto-Regressive Integrated Moving Average (ARIMA), Autocorrelation Function (ACF), Partial Autocorrelation (PACF),
binary Logistic regression, low birth weight.

I. Introduction

All the major and minor steps in life required preplanning. Using statistics, we are reviewing the datasets by applying scientific calculations for predicting the future and providing a logical view. The analysis method used to predict the future depends on the type of the dataset. The results from prediction help the industry and prepare it for future requirements. The paper is divided into two parts one is to analyze a medical dataset using binary logistic regression (Part A) for childbirth, another one is Time series new house registration (Part B) and Overseas Trip (Part C) for analyzing the time series in the annual and quarterly distribution of the time.

I. BINARY LOGISTIC REGRESSION

A. Objective

The paper analyzes the Childbirth's data set is related to health care sector for developing a binary logistic regression model for dependent variable 'lwbwt' to predict the child weight during birth using SPSS (Statistical Package for Social Sciences). The paper cab be used to help physicians [2] to prescribe the right medication and care for both the mother and infant for avoiding any health risk. Ireland has a wide sector for healthcare industry that motivates researchers to collect information and use their findings to help better decision making by physicians.

B. Introduction and data description

The data gathered from both the parents such as age, height, weight, father's education years, consumption of cigarettes. Binary logistic regression is required when the dependent variable is dichotomous (e.g.: low/high, fail/pass, yes/no). The childbirth dataset has three dichotomous variables 'smokers','lowbwt' and 'mage35'. The data set has no outliers as check using Cook's distance. The objective of the paper is to predict the low birth weight of child using binary logistic regression. So,'lwbwt' is the dependent variable.
The formula to predict the logit transformation.

$$\text{logit}(P) = \boldsymbol{a} + \boldsymbol{b}\,X$$

P is the probability of low birth weight (1) at for a given value of X, where the odd of 1 vs 0 at any point for X are P/(1-P).

The dependent variable is in fact a logit, which is a log of odds, where logit(P) is a linear function of X.

$$\textbf{Logit(P)} = \ln[P/(1-P)] = \ln(\text{odds}).$$

$$P = \frac{e^{a+bX}}{1+e^{a+bX}}$$

We can find P evaluating the coefficients **a** and **b**.

Limitations of this model is adding irrelevant variables may dilute the effects of more significant variables. The model produces more accuracy in the middle than the extremes which also predicts that may be not all the combinations exits in the sample. More number of data makes our model stable.

C. Description of the data in Childbirths

In the dataset Childbirths, total 16 variables are used with 15 independent variables length, head circumference, birthweight , ID ,Gestation ,smoker, mothers age , mnocig , mheight , mppwt , fage , fedyrs , fnocig , fheight , mage35 and one independent variable lowbwt of 42 Infants.

Name	Variable
ID	Baby number
length	Length of baby (cm)
Birthweight	Weight of baby (kg)
headcirumference	Head Circumference
Gestation **smoker**	Gestation (weeks) Mother smokes 1 = smoker 0 = non-smoker
motherage	Maternal age
mnocig	Number of cigarettes smoked per day by mother
mheight **mppwt**	Mothers height (cm) Mothers pre-pregnancy weight (kg)
fage	Father's age
fedyrs	Father's years in education
fnocig	Number of cigarettes smoked per day by father
fheight	Father's height (kg)
lowbwt	Low birth weight, 0 = No and 1 = yes
mage35	Mother over 35, 0 = No and 1 = yes

Figure 1 data description of Childbirth

Descriptive statistics is giving a clear picture of the data distribution over the population using mean, median, standard deviation of each variable. In SPSS we do it using Analysis descriptive while in R -Studio we use summary () function to get the same result.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error
Length	42	43	58	51.33	2.936	-.248	.365
Birthweight	42	1.92	4.57	3.3129	.60390	-.056	.365
Headcirc	42	30	39	34.60	2.400	.071	.365
Gestation	42	33	45	39.19	2.643	-.408	.365
smoker	42	0	1	.52	.505	-.099	.365
mage	42	18	41	25.55	5.666	.803	.365
mnocig	42	0	50	9.43	12.512	1.393	.365
mheight	42	149	181	164.45	6.504	.017	.365
mppwt	42	45	78	57.50	7.198	.492	.365
fage	42	19	46	28.90	6.864	.508	.365
fedyrs	42	10	16	13.67	2.160	-.384	.365
fnocig	42	0	50	17.19	17.308	.564	.365
fheight	42	169	200	180.50	6.978	.436	.365
lowbwt	42	0	1	.14	.354	2.118	.365
mage35	42	0	1	.10	.297	2.861	.365
Valid N (listwise)	42						

Figure 2 Summary statistics of childbirths

Here, in Childbirths we have total number of observation for length variable number of observations 42(N=42).Minimum length of a infant during birth is 43 cm and maximum 58cm with an average length 51.33cm and variance as 2.936cm.Birthweight of a child shows 1.92kg is the minimum weight a child is born with and maximum 4.57kg with a average of 3.31kg and very slight variation with 0.60kg means most of the child in the dataset are underweight which contributed to our study where we analyze the factors responsible for low birth weight of Infant in Ireland. The side of the head is minimum 30 and maximum 39 which is one of our responsible variables for low birth weight of baby. Various studies are done to correlate the head circumference of a baby to its weight using the two-tailed tests and Chi- square tests. In most of the cases as per various studies in medical field the low birth is associated with mother's old age as the minimum age 18year and maximum 41 years in childbirths dataset gives a inference that an average age of women to get pregnant is 25 years. Smoking habit also leads to low birth weight as it impacts the lungs [3] directly and causes health issues in later life to the infant as well. We have smoker as a dichotomous variable that will be combined with other factors to predict the low birth weight of an Infant.

Frequency table shows there is no missing values in the dataset and describe the data distribution of each variable. Few of the variable used in models are shown in the frequency table and histogram for visual representation below.

lowbwt

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	36	85.7	85.7	85.7
	1	6	14.3	14.3	100.0
	Total	42	100.0	100.0	

Figure 3 frequency distribution of low birth rate

mage35

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	38	90.5	90.5	90.5
	1	4	9.5	9.5	100.0
	Total	42	100.0	100.0	

Figure 4 frequency distribution of mother's age

The graphical representation shows that most of mother's smokes less than 10 cigarettes. The age of mother's mostly lies between 20-23 years.

smoker

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	20	47.6	47.6	47.6
	1	22	52.4	52.4	100.0
	Total	42	100.0	100.0	

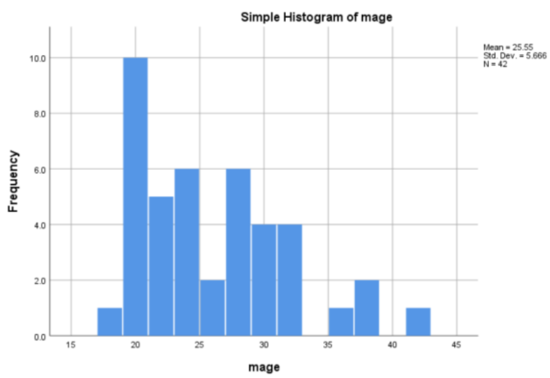


Figure 5 Graphical representation of the mother age

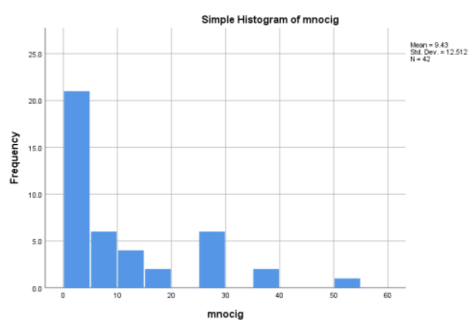


Figure 6 Histogram representing number of cigarettes.

Binary logistic regression must go through four assumptions:

i. The dependent variable should be dichotomous example low birth has two values 0 or 1 it can be used.
ii. One or more than one independent variable should be there. Example we have mother age, mothers' weight before pregnancy, whether mother smokes or not.
iii. Independence of observation needs to be present where the dependent variable.
iv. There needs to be a linear relationship between the independent and the logit transformed dependent variable.

Coefficients[a]

Model		Collinearity Statistics Tolerance	VIF
1	ID	.652	1.534
	Length	.241	4.155
	Birthweight	.195	5.128
	Headcirc	.401	2.494
	Gestation	.295	3.390
	smoker	.343	2.918
	mage	.103	9.740
	mnocig	.365	2.739
	mheight	.311	3.220
	mppwt	.405	2.472
	fage	.147	6.791
	fedyrs	.578	1.731
	fnocig	.534	1.872
	fheight	.540	1.850
	mage35	.249	4.015

a. Dependent Variable: lowbwt

Figure 7collinearity statistics

Models feature selection usually done by checking the collinearity where the motive of the study is to get model fit without overfitting the features and values less than 0.2 and VIF close to 10.

D. Model Building and Evaluation

Logistic regression produces two models with the set of variables that are added in model 1 but not in Model 0 which is called null model. Model fit can be checked by the Wald statistics for each variables contribution in model by using statistical significance using chi-square. The dataset has small values N=42 that results in smaller significance of Chi-square test. Most of the cases by including all the variables that contributed to low birth rate we attain more than 90% accuracy that is misleading as it may not work on new data. So, we check the percentage of cases we assume in both the models using classification report generated from both the models. Models can be unstable in small samples.

Baseline Model: The overall accuracy of the null model 85.7 %. Which shows that in the sample most of the babies are not underweight.

Block 0: Beginning Block

Classification Table[a,b]

Observed			Predicted lowbwt 0	Predicted lowbwt 1	Percentage Correct
Step 0	lowbwt	0	36	0	100.0
		1	6	0	.0
	Overall Percentage				85.7

a. Constant is included in the model.

b. The cut value is .500

Figure 8 Null model of childbirths

Model fixing by rejecting the intermediate model is done as accuracy comparison is done by changing the independent variables impact on the dependent variable. By using the head circumference of the child, the low birth weight cannot be predicted accurately as the model is giving only 38.4% accuracy.

Model 1: Independent variable used is 'headcirc' Head circumference of the baby.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	24.284[a]	.215	.384

Figure 9 Model building using independent variable' headcirc'

Classification Table[a]

Observed			Predicted lowbwt 0	Predicted lowbwt 1	Percentage Correct
Step 1	lowbwt	0	35	1	97.2
		1	4	2	33.3
	Overall Percentage				88.1

a. The cut value is .500

Figure 10 Classification model 1

Model 2: Predict the accuracy 60.5% of low birth weight of a infant when the independent variables mothers age, mother weight before pregnancy, mother age above 35 or not, head circumference of the baby is included.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	17.098[a]	.338	.605

Figure 11 logistic regression model 2

Classification Table[a]

Observed			Predicted lowbwt 0	Predicted lowbwt 1	Percentage Correct
Step 1	lowbwt	0	36	0	100.0
		1	2	4	66.7
	Overall Percentage				95.2

a. The cut value is .500

Figure 12 classification table of model 2

Model 3: Predict the accuracy 74.5% which is a good model, but we need to find the model that has more accuracy than the null model.so rejecting this model 3 also.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	11.771[a]	.417	.745

Figure 13 logistic regression model 3

Classification Table[a]

Observed			Predicted lowbwt 0	Predicted lowbwt 1	Percentage Correct
Step 1	lowbwt	0	35	1	97.2
		1	1	5	83.3
	Overall Percentage				95.2

a. The cut value is .500

Figure 14 classification table of model 3

Model 4: Predict the accuracy 100% when more independent variable is used but the model, we design using the less variable with high accuracy need to be evaluated. More fitted model need not to be included as it may not work for new data.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	.000[a]	.560	1.000

Variable(s) entered on step 1: mppwt, mage, Headcirc, mage35, Length, Birthweight, Gestation, smoker, mnocig, mheight, fage, fedyrs, fnocig, fheight.

Classification Table[a]

Observed			Predicted lowbwt 0	Predicted lowbwt 1	Percentage Correct
Step 1	lowbwt	0	36	0	100.0
		1	0	6	100.0
	Overall Percentage				100.0

a. The cut value is .500

We removed the overfitted, underfitted models and choose the below model with 97.3% accuracy which predicts the low birth weight of a child.

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	24.200	5	.000
	Block	24.200	5	.000
	Model	24.200	5	.000

Figure 15 final logistic regression model

The model summary shows that no missing data.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	10.250[a]	.438	.783

a. Estimation terminated at iteration number 9 because parameter estimates changed by less than .001.

Figure 16 highly accurate model

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	3.913	8	.865

Figure 17 Hosmer and Lemeshow test for final model

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	42	100.0
	Missing Cases	0	.0
	Total	42	100.0
Unselected Cases		0	.0
Total		42	100.0

Figure 18summary of final model in childbirth

Classification Table[a]

	Observed		Predicted lowbwt 0	Predicted lowbwt 1	Percentage Correct
Step 1	lowbwt	0	36	0	100.0
		1	1	5	83.3
	Overall Percentage				97.6

a. The cut value is .500

Figure 19 Classification table of final model

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	mage35(1)	-4.607	5.236	.774	1	.379	.010
	mppwt	-.356	.225	2.514	1	.113	.700
	mnocig	.066	.076	.754	1	.385	1.068
	Gestation	-1.268	.606	4.379	1	.036	.282
	mage	-.137	.272	.253	1	.615	.872
	Constant	71.969	31.288	5.291	1	.021	1.802E+31

a. Variable(s) entered on step 1: mage35, mppwt, mnocig, Gestation, mage.

The p value is .000 <.005 which shows the good fit of a model and it is accepted to predict the results.
The cox n Snell R square(pseudo R) explain variation in the dependent variable for this model range from 43.8% to 78.3%.

E. Evaluation and Conclusion from Models

Model	-2log likelihood	Nagelkerke P square	Accuracy
1	24.284	.384	88.1
2	17.098	.605	95.2
3	11.771	.745	95.2
4	.000	1.000	100
5	6.575	.867	.783
6	10.250	.783	97.3

Figure 20 Accuracy of different models

The last model (6) with independent variables 'mage','Gestation','mppwt' and,'mnocig' has highest accuracy which is 97.6% with less value of -2loglikehood (10.250) which shows least deviation in the data. Also, the p value is less than 0. 005.Both the tests Omni bus and Hosmer and Lemeshow are giving significant values The shows that the model fit with the chosen variables on dependent variable low birth weight of a child.
Resultant equation for the model with highest accuracy is:
Ln(odds)= 71.96+0.66*mnocig-0.35*mppwt-4.60*mage35-0.13*mage-1.26*Gestation

II. TIME SERIES -PART B (OVERSEAS TRIPS)

A. Objective

The dataset Oversea trip aim to predict the count of the travelers in Ireland using the historical quarterly data from 2012(Quarter 1) to 2019 (Quarter 4).

B. Introduction

The dataset has two variables quarterly and trips in thousands that is distributed among 32 observations recorded from previous year and we need to predict the next three quarters for the same. Using the R- Error, S-seasonality (as it has quarterly data), T- trend, C- cyclic pattern.

	Qtr1	Qtr2	Qtr3	Qtr4
2012	1165.1	1817.3	2096.7	1438.0
2013	1251.7	1893.0	2261.0	1580.1
2014	1342.5	2126.6	2440.4	1694.9
2015	1531.3	2344.9	2770.9	1995.9
2016	1784.7	2598.9	3061.5	2139.2
2017	1796.1	2769.4	3095.6	2270.9
2018	1920.7	2951.9	3330.9	2412.8
2019	2026.7	3021.8	3334.4	2424.6

Figure 21 Quarterly trips from 2012 to 2019 to Ireland

Analysis of the oversea trip needs to be done by checking the pattern using visualization it produces in plot () function that indicates its shape.

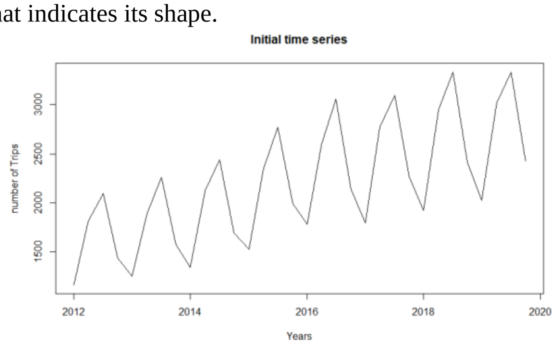


Figure 22 Overseas trip historic quarterly data

Time series exhibit the trend from its past patterns that helps in future predict. Various businesses such as Airline, Customer Services, Tour, and travel work using time series only to sustain the market. The graph shows that there is a trend (linearly increasing upward) and seasonality (peaks) factor involved in time series.

The maximum number of tourists[1] in Ireland are coming in third quarter.

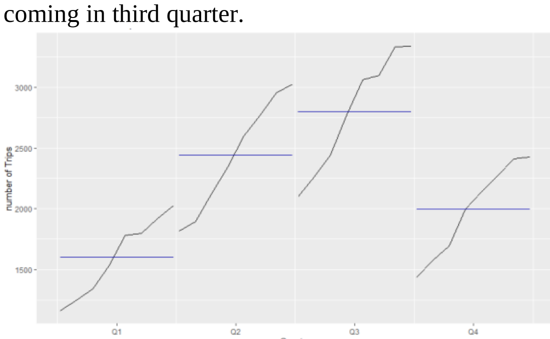


Figure 23 quarter 3 has maximum tourists.

The pattern requires to be smoothened by eliminate noises using ma (moving average) that uses mean value.

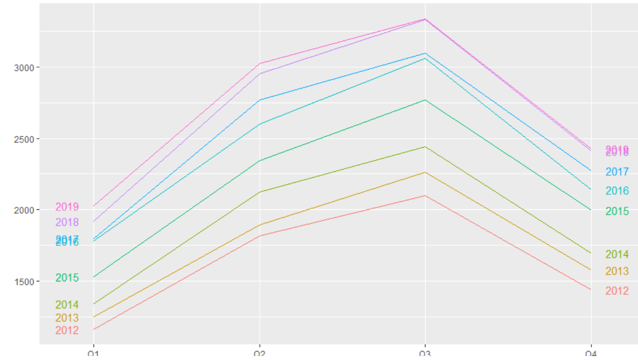


Figure 24 seasonal plot overseas trip

Using the k value (by losing (k-1)/2 observations at each end) as we increase k the graph pattern become smooth. There are two type of seasonal decomposition additive and multiplicative.

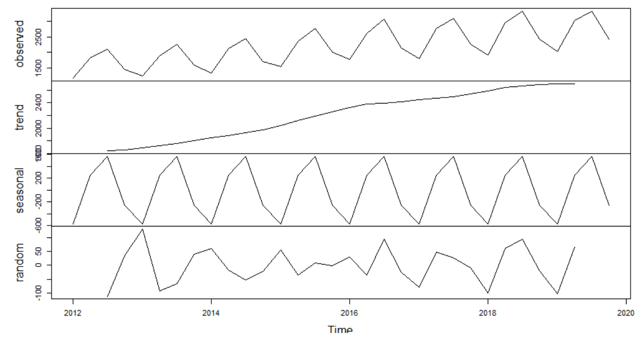


Figure 25 additive seasonality decomposition overseas

Additive decomposition can be calculated using the observation time t for seasonality time t and irregularity time t.

Yt= Trendt+ Seasonalt+ Irregulart

Multiplicative decomposition

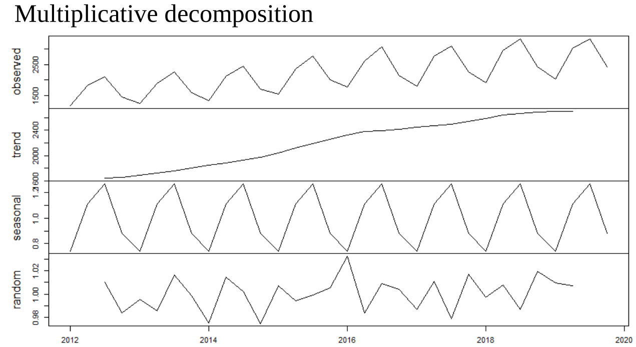


Figure 26 multiplicative seasonality decomposition overseas trips

Yt= Trendt* Seasonalt* Irregulart

C. MODEL BUILDING

Model 1 : Naïve model is predicting the next three quarters using the historic pattern.The Accuracy is 638.5391 of the training set in the oversea trip using the naïve model.

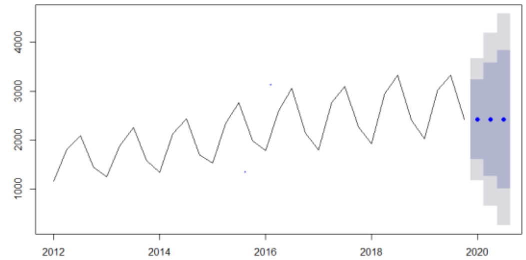


Figure 27 Naive model overseas trip

```
                 ME     RMSE      MAE      MPE     MAPE     MASE        AC
Training set 40.62903 638.5391 576.9581 -1.88058 26.39067 3.765343 -0.011388

Forecasts:
        Point Forecast    Lo 80    Hi 80     Lo 95    Hi 95
2020 Q1         2424.6 1606.279 3242.921 1173.0863 3676.114
2020 Q2         2424.6 1267.320 3581.880  654.6923 4194.508
2020 Q3         2424.6 1007.227 3841.973  256.9146 4592.285
```

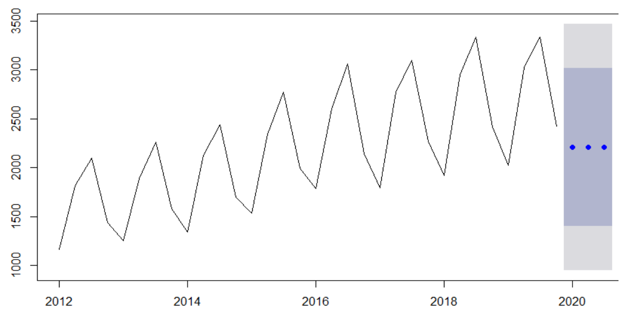


Figure 28 Mean model for oversea trip.

Model 2: Mean model is the most basic model that we use to analyze in statistics using the independent and similar values distribute over the dataset.

```
                      ME    RMSE      MAE       MPE     MAPE     MASE
Training set 1.98952e-13 598.416 497.1109 -8.180872 24.99967 3.24424

Forecasts:
        Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
2020 Q1       2209.075 1400.589 3017.561 949.8425 3468.308
2020 Q2       2209.075 1400.589 3017.561 949.8425 3468.308
2020 Q3       2209.075 1400.589 3017.561 949.8425 3468.308
> plot(fcast.mean)
```

The RMSE for mean model is 598.416 for the predicted quarters in 2020 (Q1, Q2,Q3).

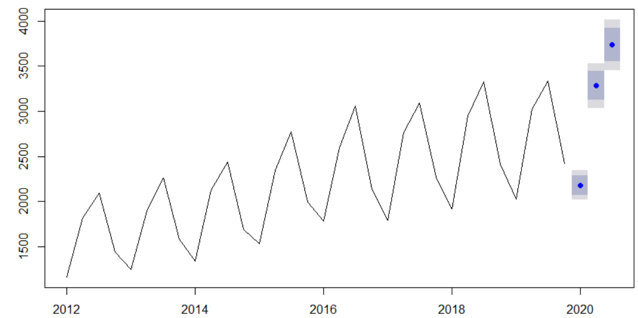


Figure 29 holts winter model for overseas trip

```
                   ME     RMSE      MAE       MPE     MAPE      MASE      ACF1
Training set -16.59663 72.44923 57.49647 -1.052437 2.753433 0.3752334 0.6343327

Forecasts:
        Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
2020 Q1       2180.207 2072.930 2287.483 2016.141 2344.272
2020 Q2       3284.157 3122.532 3445.782 3036.973 3531.342
2020 Q3       3734.601 3550.769 3918.432 3453.454 4015.747
> plot(fcast.hw)
```

Model 3: Holt (1957) and Winters(1960) evolve this method for predicting the seasonality in the time series using the three smoothing methods for trend, seasonal component and smoothing. The RMSE of holt's winter model is 72.44 indicating the next three quarters using the three dots at the right of the graph for the year 2020.

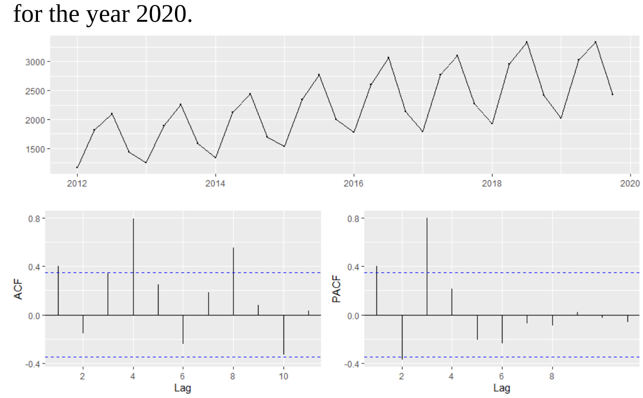


Figure 30 check the difference for ACF and PACF lag

ACF is the auto-correlation function for time series it basically shows that how the future values of time series is related to the past values using the trend, seasonality ,cyclic and residual values the complete auto correlation graph is produced.

PACF is partial auto-correlation function which has lags to find correlation of residual with the next lag to find next correlation.

Model 4: ARIMA, this model is also called autoregressive moving average model. It is used to study the time series for predicting the future trends.

Residual Testing

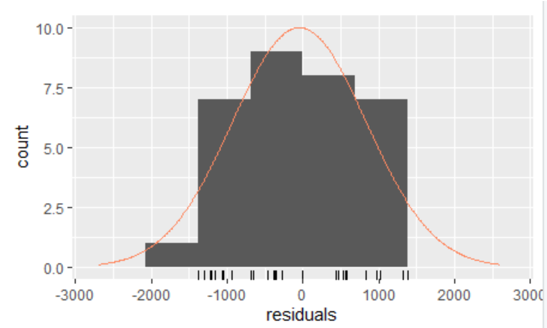


Figure 31 residual graph from ARIMA

```
        Ljung-Box test

data:  Residuals from ARIMA(0,2,0)
Q* = 86.711, df = 6, p-value < 2.2e-16

Model df: 0.   Total lags used: 6
```

```
                   ME     RMSE      MAE       MPE     MAPE     MASE      ACF1
Training set -48.79949 869.383 772.4571 -2.278997 35.81204 5.041208 0.0105789
```

ARIMA model is showing the RMSE 869.383 with a quarterly forecast.

```
        Point Forecast        Lo 80     Hi 80       Lo 95     Hi 95
2020 Q1         1514.8    364.1014 2665.499   -245.0417 3274.642
2020 Q2          605.0  -1968.0403 3178.040  -3330.1256 4540.126
2020 Q3         -304.8  -4610.3200 4000.720  -6889.5246 6279.925
> plot(fcast)
Series: tost
ARIMA(1,0,0)(0,1,0)[4] with drift

Coefficients:
         ar1    drift
      0.5835  35.9414
s.e.  0.1585   7.9346

sigma^2 estimated as 5616:  log likelihood=-159.77
AIC=325.53   AICc=326.53   BIC=329.53
```

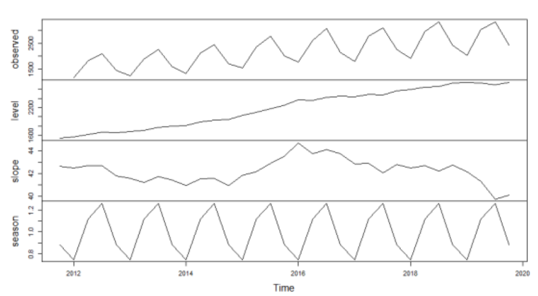


Figure 32 ETS model for overseas trip

```
> accuracy(fcast.auto)
                   ME     RMSE      MAE        MPE     MAPE      MASE     -0.0
Training set -7.22914 54.69822 44.54334 -0.3013343 2.017642 0.2906987 -0.0
```

The ETS model has the RMSE 72.44 for the overseas trip.

D. Evaluation and Conclusion in Overseas Trip

The study on the dataset timeseries forecast the next three quarters accuracy using the RMSE(Root mean square metric) by decomposing the raw time series that had noise and models such as Holt-winters predicting RMSE value 72.449,ETS model has 72.44 with other models such as automated ARIMA outperformed with RMSE 869.383, mean forecast , Naïve model prediction of trend and seasonality is plotted in the graphs for next three quarters in 2020 after the 2019 quarter 4.

```
> accuracy(fit)
                   ME     RMSE      MAE       MPE     MAPE     MASE      ACF1
Training set -48.79949 869.383 772.4571 -2.278997 35.81204 5.041208 0.0105789
> accuracy(fcast.ses)
                  ME     RMSE      MAE      MPE     MAPE     MASE       ACF1
Training set 147.9911 521.9777 473.5028 1.888505 21.50027 3.090173 -0.0133416
> accuracy(hfit2)
                  ME     RMSE      MAE      MPE     MAPE     MASE        ACF1
Training set 147.9183 521.9777 473.5527 1.885681 21.50299 3.090499 -0.01339834
> accuracy(fcast.auto)
                  ME     RMSE      MAE        MPE     MAPE      MASE        ACF1
Training set -7.22914 54.69822 44.54334 -0.3013343 2.017642 0.2906987 -0.05187753
> accuracy(fcast.mean)
Error in accuracy.default(fcast.mean) :
  Unable to compute forecast accuracy measures
> accuracy(fcast.naive)
                  ME     RMSE      MAE      MPE     MAPE     MASE        ACF1
Training set 40.62903 638.5391 576.9581 -1.88058 26.39067 3.765343 -0.01138884
> accuracy(fcast.hw)
                   ME     RMSE      MAE       MPE     MAPE      MASE      ACF1
Training set -16.59663 72.44923 57.49647 -1.052437 2.753433 0.3752334 0.6343327
```

III. TIME SERIES -PART B (NEW HOUSE REGISTRATION)

A. Objective

The objective of this paper is to analyze the number of houses registered in Ireland from 1978 to 2019 year using the time series . The data has annual frequency with two variables 'Years' and 'Newhouseregistration'.It has 42 records of the houses each record indicating the number of houses registered in that year in Ireland.

B. Introduction

The dataset is from CSO.ie website which and analyzing this dataset. Using the R-Studio the time series analysis completed

```
setwd("~/DMML/New folder/project/")
library(fpp2)
library(tseries)
df<-read.csv(file='House.csv',col.name=c("Year","NewHouseRegistrations"))
df
House<-ts(df$NewHouseRegistrations,start= 1978,frequency=1)
plot(House,main="Time series House",xlab="Years",ylab="number of houses registered")
summary(House)
plot(House)
```

It helps the real estate agents as well to check the market trend to employee new people in their firms as per the predicted values for next three years. The below graph indicates the rise and fall in registrations between the time 2000 to 2010 in Ireland. The minimum 627 houses were registered in a year and maximum 66649 houses where the maximum registrations are done in the third quarter which is predicted using the past data available in dataset in summary function.

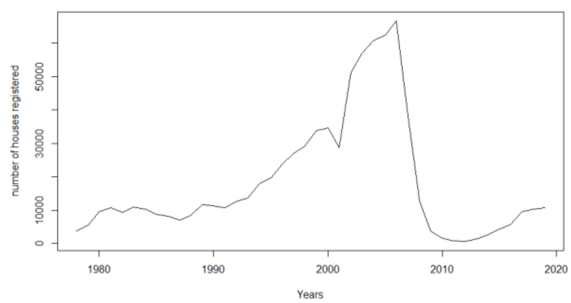


Figure 33 Raw dataset distributed over time series graph.

The rise and fall are noises in time series which can be handled by using the appropriate smoothing techniques such as ma (moving average) function and it can be analyzed using the models of time series.

C. Model building and Evaluation

Model 1: ARIMA model, autoregressive moving average model uses the three components p-auto regressive portion of PACF, d- difference and q – moving average portion from ACF plot.The adf (Augmented dickey fuller test shows that the value is not less than 0.05 and we need to consider the differencing (difhouse<-diff (House))which can normalize the mean by eliminating the trend .

```
        Augmented Dickey-Fuller Test

data:  difhouse
Dickey-Fuller = -2.6751, Lag order = 3, p-value = 0.308
alternative hypothesis: stationary
```

ACF eliminate the correlation to avoid multicollinearity.

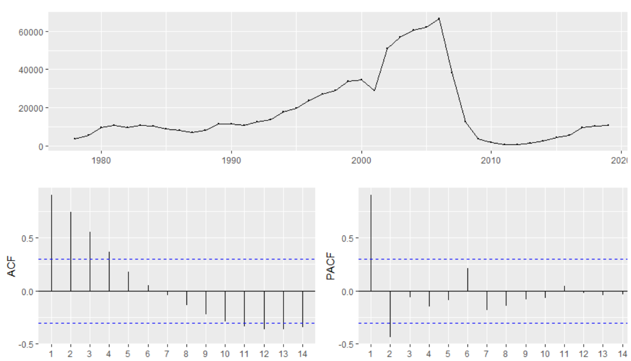


Figure 34 ACF and PACF graph

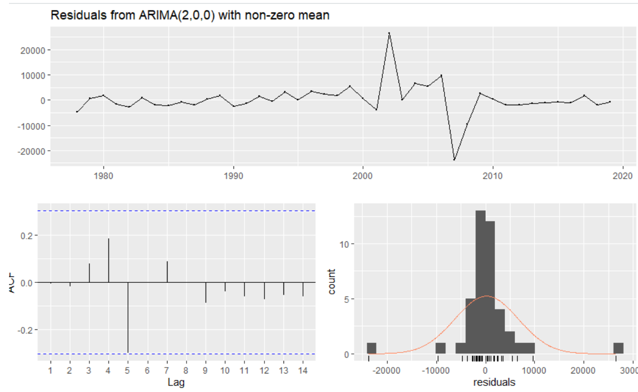


Figure 35 Residual graph in ARIMA Model

```
Series: House
ARIMA(2,0,0) with non-zero mean

Coefficients:
         ar1      ar2       mean
      1.3346  -0.4665  16791.106
s.e.  0.1315   0.1319   6985.186

sigma^2 estimated as 43317727:  log likelihood=-428.43
AIC=864.86   AICc=865.94   BIC=871.81
> checkresiduals(fit_arima)

        Ljung-Box test

data:  Residuals from ARIMA(2,0,0) with non-zero mean
Q* = 6.8201, df = 5, p-value = 0.2344

Model df: 3.   Total lags used: 8
```

The Q-Q plot testing shows the normal distribution of data over the period is present in the dataset. The fitted models forecast the next three years. Which predicts that the first year ,second year ,third year high ,low,average registrations of new houses.

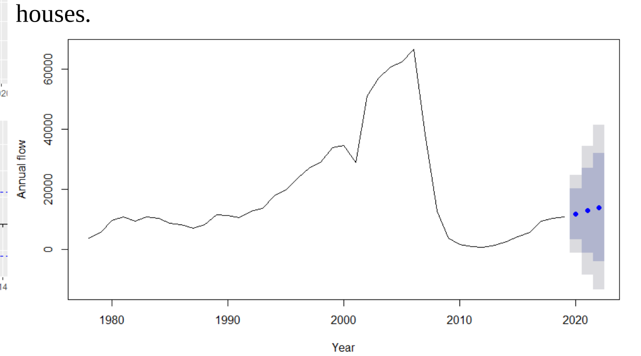


Figure 36 Prediction for ARIMA

```
     Point Forecast     Lo 80    Hi 80       Lo 95    Hi 95
2020       11818.59  3383.902 20253.27   -1081.151 24718.33
2021       12957.23 -1109.161 27023.62   -8555.457 34469.91
2022       13994.20 -3917.264 31905.66  -13399.019 41387.42
```

Model 2: Basic Mean model

```
Error measures:
                       ME     RMSE      MAE       MPE     MAPE    MASE      ACF1
Training set 1.559007e-12 17881.98 14062.65 -241.9622 271.8443 3.54469 0.9049882

Forecasts:
     Point Forecast     Lo 80    Hi 80     Lo 95    Hi 95
2020       18275.33 -5578.056 42128.72 -18708.38 55259.05
2021       18275.33 -5578.056 42128.72 -18708.38 55259.05
2022       18275.33 -5578.056 42128.72 -18708.38 55259.05
```

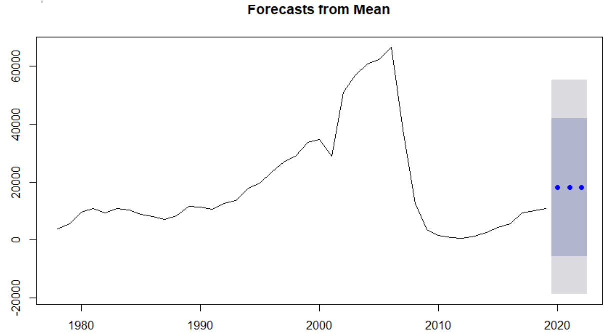
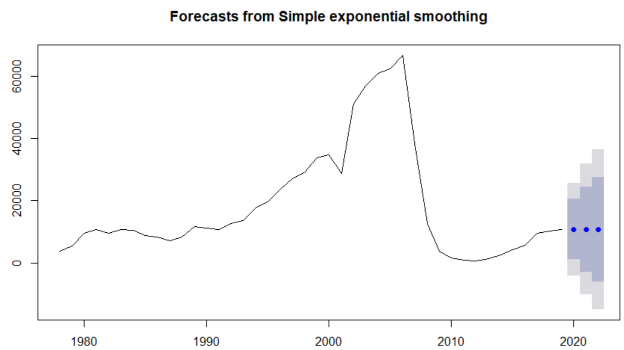


Figure 37 Prediction plot of mean model

The model shows exact number for all the three consecutive years house registration prediction the accuracy of the model will be checked which is outperformed as 17881.98.

Model 3: Simple exponential smoothing model.

The future trend is shown by the grey area in graph for the model. The Root mean square exponential is 7413.074 for the below model.




```
#Automated ETS()
fcast.auto<-ets(House,model="ZZZ")
plot(fcast.auto)
summary(cast.auto)
accuracy(fcast.auto)
forecast(fcast.auto,h=3)

#mean model
fcast.mean<-meanf(House,h=3)
summary(fcast.mean)
plot(fcast.mean)

#naive model
fcast.naive<-naive(House,h=3)
summary(fcast.naive)
plot(fcast.naive)


#seasonal naive model
fcast.snaive<-snaive(House,h=3)
summary(fcast.snaive)
plot(fcast.snaive)
```

Figure 38 model building and evaluation

```
> accuracy(fit_arima)
                   ME     RMSE      MAE       MPE     MAPE      MASE         ACF1
Training set 207.1252 6342.208 3464.418 -20.20197 35.95662 0.8732557 -0.007018081
> accuracy(fcast.ses)
                   ME     RMSE      MAE       MPE     MAPE      MASE      ACF1
Training set 165.2083 7378.822 3875.789 -7.771767 33.33161 0.9769476 0.4218681
> accuracy(sfit2)
                   ME     RMSE      MAE       MPE     MAPE     MASE      ACF1
Training set 55.00544 7413.074 3984.802 -10.67123 36.22512 1.004426 0.4141101
> accuracy(sfit1)
                   ME     RMSE      MAE       MPE     MAPE     MASE      ACF1
Training set 55.00544 7413.074 3984.802 -10.67123 36.22512 1.004426 0.4141101
> accuracy(fcast.auto)
                    ME     RMSE      MAE       MPE     MAPE     MASE      ACF1
Training set -184.0889 7395.984 3870.015 -14.83976 36.43003 0.975492 0.4173692
> accuracy(fcast.mean)
                       ME     RMSE      MAE       MPE     MAPE    MASE      ACF1
Training set 1.559007e-12 17881.98 14062.65 -241.9622 271.8443 3.54469 0.9049882
> accuracy(fcast.snaive)
                   ME     RMSE      MAE       MPE     MAPE MASE      ACF1
Training set 170.8049 7466.737 3967.244 -7.900382 34.07645    1 0.4216754
> accuracy(fcast.naive)
                   ME     RMSE      MAE       MPE     MAPE MASE      ACF1
Training set 170.8049 7466.737 3967.244 -7.900382 34.07645    1 0.4216754
> mage
```

Figure 39 Accuracy metrices of models in dataset

D. Conclusion:

Data Analysis on the ‘newhouseregistration’ concludes that the basic models with insignificant RMSE values were excluded from the paper as the resultant models are shown in Figure31. the new house registration prediction for 2020,2021,2022 completed by evaluating the RMSE, AIC metrics using the graphs generated by models with RMSE evaluated for ARIMA(7342.208) which satisfied Box-Ljung test of model fitting, Simple exponential (7378.822) ,auto regression (7395.984) the models by fitting into the dataset and decomposing the noises.

REFERENCES

[1] Choden, & Unhapipat, Suntaree. (2018). ARIMA model to forecast international tourist visit in Bumthang, Bhutan. Journal of Physics: Conference Series. 1039. 012023. 10.1088/1742-6596/1039/1/012023.

[2] Desalegn Dargaso Dana, Binary Logistic Regression Analysis of Identifying Demographic, Socioeconomic, and Cultural Factors that Affect Fertility Among Women of Childbearing Age in Ethiopia, Science Journal of Applied Mathematics and Statistics. Vol. 6, No. 3, 2018, pp. 65-73. doi: 10.11648/j.sjams.20180603.11

[3] Currie, J., Neidell, M. (2005). Air Pollution and Infant Health: What Can We Learn from California's Recent Experience? The Quarterly Journal of Economics (2005)120(3): 1003-1030. doi: 10.1093/qje/120.3.1003.