Statistics CA2: Time series analysis and Logistic Regression

1st Dhwani Dharmesh hingu
Dept. of Computing
National College of Ireland
Dublin, Ireland
x19216742@ncirl.student.ie

***Abstract*—This paper demonstrates time series analysis and classification in statistics using three datasets. In Part A, OverseasTrips and NewHouseRegistrations datasets are used, both of these datasets consist of time series data. "OverseasTrips" is a quarterly time series of non-residents visiting Ireland from the first quarter of 2012 to the fourth quarter of 2019 whereas from 1978 to 2019, 'NewHouseRegistrations' is an annual collection of new house registrations. On OverseasTrips datasets average mean method, naïve method and arima model is used and on NewHouseRegistrations average mean method, seasonal naïve method and sarima (seasonal arima) model is used. Whereas, for Part B Childbirths dataset logistic classification is done using SPSS and was able to get 97.6 accuracy.**

I. Part A

A. Dataset 1: NewHouseRegistrations

1) Data Description and Data Understanding: This data set is about new house registrations from 1978 to 2019, the dataset is an annual time series. The datasets are from the data repository of the Central Statistics Office in Ireland. In this dataset the very first thing was done to make the data time series data using ts(data, start = 1978, frequency = 1) function in R, where the start is 1978, end is 2019 as it is given in data description and frequency had to be 1 as this is annual series data. Later on, graph was plotted to check the pattern of the data. By fig.1 it's been observed that the data has a trending pattern as by every year there is an uptrend followed in linear fashion as well as there is a dip to be noticed between 2000 to 2010 and again the pattern is observed and that it is growing again. Thus, it was concluded that the data was trending series and there were no seasonal patterns as well. The next step was to check whether the data is stationary or not, so ndiffs() function was used and got to know that the data was stationary as it returned 0.

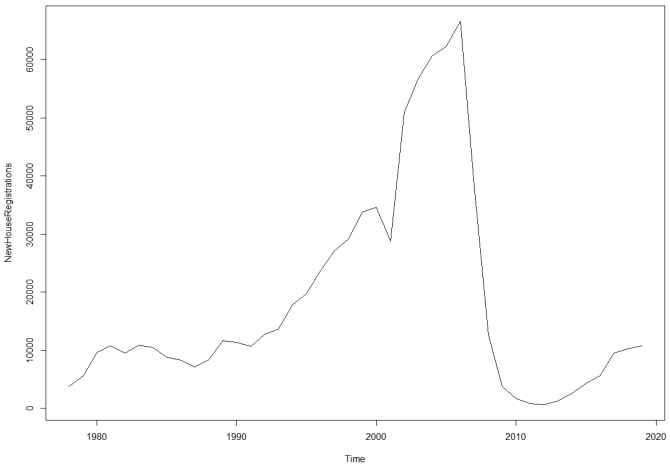


Fig. 1. Plot for NewHouseRegistration Data

2) Models:

- Average mean: In model building the simplest model simple average mean was executed using meanf(ts_NewHouseReg, h=3) function as we want to forecast future 3 data points. In fig.2 and fig.3 it shows that the model forecasted 3 observations which is "18275" for 2020,2021 and 2022. Once the values were forecasted the model was checked against residual plot, RMSE and p-value, so it was noticed that the RMSE was 17881 which is very high and acceptable, p-value was 6.661e-16 which is not greater than 0.05. Thus, dropped this model and looked for naïve mode.
- Naïve(Random walk): It is the most basic type of model. According to this method, the forecast for any time equals the actual value of the previous period. Since only one last cycle of data is needed for forecasting, this model does not require a large number of data points. Seasonality, pattern, or both may be considered [1]. The data simply set all forecasts to the value of the last observation for naive forecasts. The model is build using naive(ts_NewHouseReg) function and it forecasted three observations based on the former data observations is. As seen in the fig.6 the RMSE score was 7466 and its still high and not acceptable although it got decreased compared to average mean model and p-value was 0.04 which is nor greater than 0.05. Residual plot was also plotted to check whether the lags are inside the significant boundary or no and it was observed that the lags were going outside. However, this model was also dropped and simple exponential was built to check to get best model.

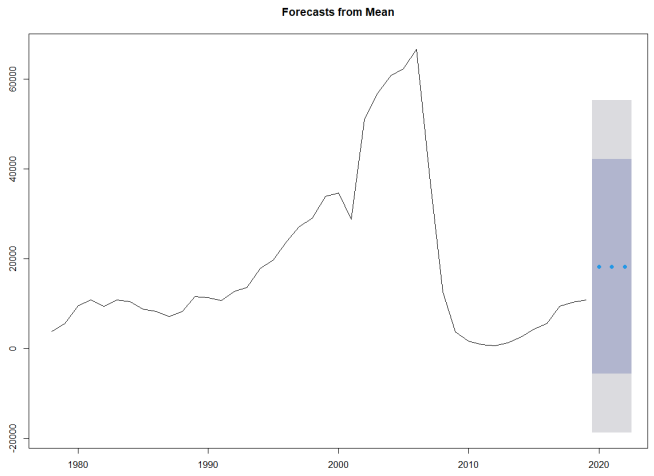


Fig. 2. Forecast Plot of Average Mean Model for NewHouseReg Data

```
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

        Ljung-Box test

data:  Residuals from Mean
Q* = 86.423, df = 7, p-value = 6.661e-16

Model df: 1.   Total lags used: 8
```

Fig. 3. Summary of Average Mean Model for NewHouseReg Data

- Simple exponential: Simple exponential is appropriate and is used when the data is trending, is not seasonal and an additive method with constant level is witnessed. In this model the RMSE as shown in fig.9 got little decreased and p-values is 0.01 which is not greater than 0.05 as well as, as per the residual plot the lags are going out of the significant boundary in ACF. This model was dropped as well and Holt Exponential was executed to get the best model.

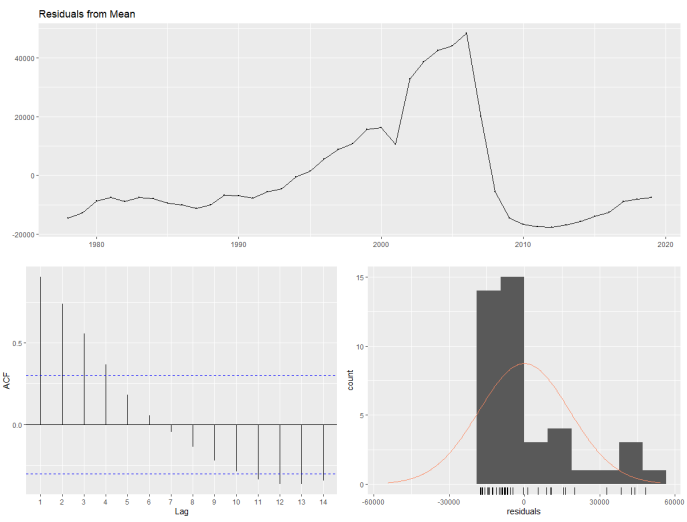


Fig. 4. Residual Plot of Average Mean Model for NewHouseReg Data

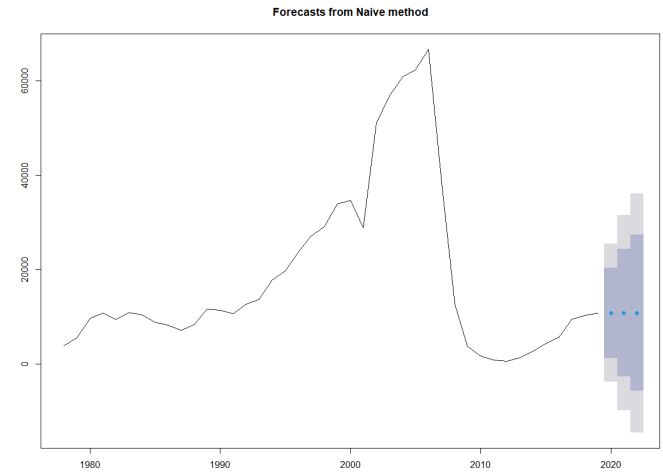


Fig. 5. Forecast Plot of Naive Model for NewHouseReg Data

```
Error measures:
                    ME     RMSE      MAE       MPE     MAPE MASE      ACF1
Training set 170.8049 7466.737 3967.244 -7.900382 34.07645    1 0.4216754

Forecasts:
     Point Forecast     Lo 80    Hi 80     Lo 95    Hi 95
2020          10784  1214.992 20353.01 -3850.535 25418.54
2021          10784 -2748.621 24316.62 -9912.358 31480.36
2022          10784 -5790.009 27358.01 -14563.759 36131.76

        Ljung-Box test

data:  Residuals from Naive method
Q* = 15.919, df = 8, p-value = 0.04356

Model df: 0.   Total lags used: 8
```

Fig. 6. Summary of Naive Model for NewHouseReg Data

- Holt Exponential:Holt Exponential is appropriate and used when the data is trending, does not have seasonality and additive model with uptrend or downtrend is witnessed. The model was built using holt(ts_NewHouseReg) function and a forecast plot was plotted to get next three forecasted and values. “11874.53”, “12964.94 ” and “14055.35” was forecasted for 2020, 2021 and 2022 as seen in the fig.11 & fig.12

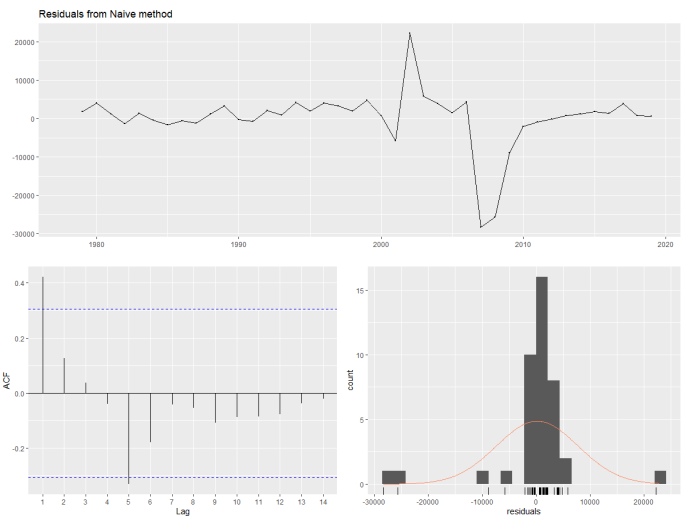


Fig. 7. Residual Plot of Naive Model for NewHouseReg Data

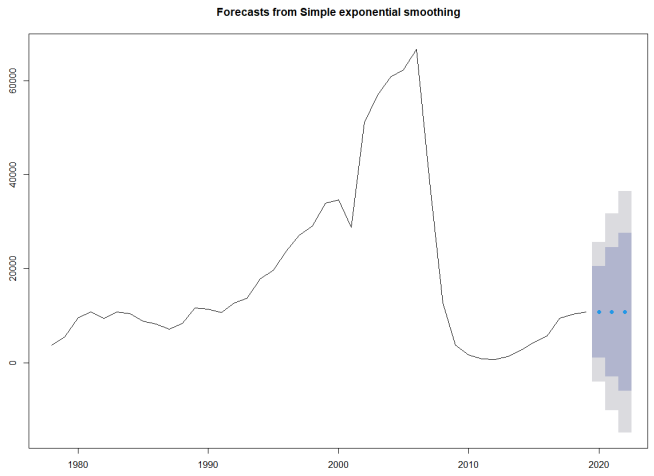
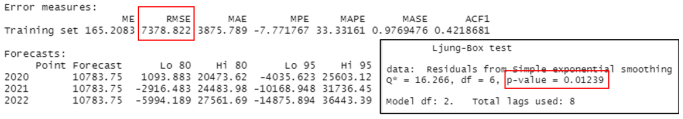


Fig. 8. Forecast Plot of Simple exponential for NewHouseReg Data



```
Error measures:
                    ME    RMSE      MAE       MPE     MAPE      MASE      ACF1
Training set 165.2083 7378.822 3875.789 -7.771767 33.33161 0.9769476 0.4218681

Forecasts:
     Point Forecast     Lo 80    Hi 80     Lo 95    Hi 95
2020       10783.75  1093.883 20473.62  -4035.623 25603.12
2021       10783.75 -2916.483 24483.98 -10168.948 31736.45
2022       10783.75 -5994.189 27561.69 -14875.894 36443.39

        Ljung-Box test

data:  Residuals from Simple exponential smoothing
Q* = 16.266, df = 6, p-value = 0.01239

Model df: 2.   Total lags used: 8
```

Fig. 9. Summary of Simple exponential for NewHouseReg Data

.It was noticed that RMSE got increased compared to simple exponential model and p-value got little better but still was that good enough. There was not much different in simple exponential and holt exponential the only improvement was seen was in residual plot where the lags were less outside the significant boundary. This model was kept on hold and another model was built using ARIMA to get the best model.

- Autoregressive Integrated Moving Average

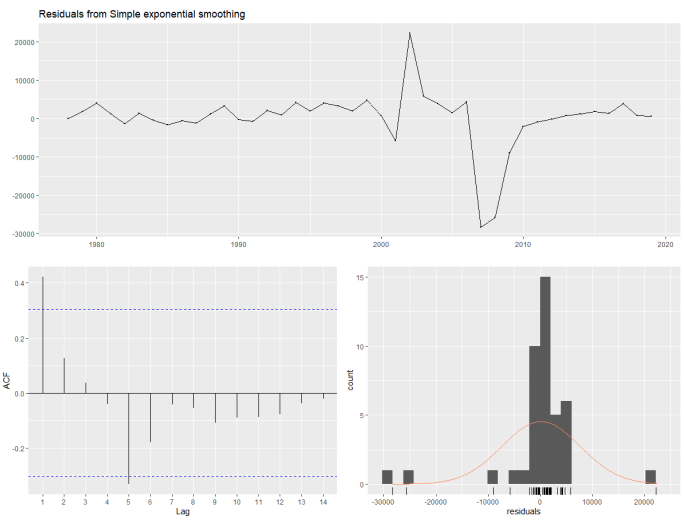


Fig. 10. Residual Plot of Simple exponential for NewHouseReg Data

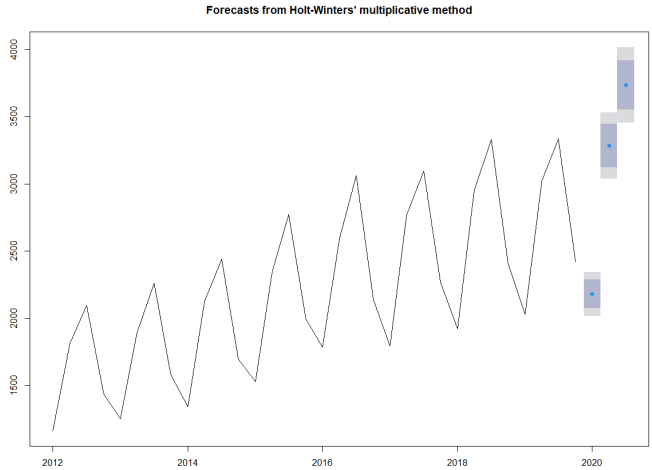
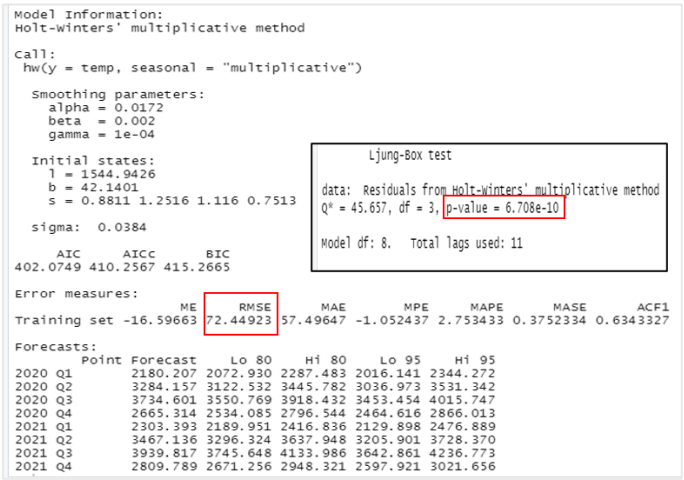


Fig. 11. Forecast Plot of Holt Exponential for NewHouseReg Data



```
Model Information:
Holt-Winters' multiplicative method

Call:
 hw(y = temp, seasonal = "multiplicative")

  Smoothing parameters:
    alpha = 0.0172
    beta  = 0.002
    gamma = 1e-04

  Initial states:
    l = 1544.9426
    b = 42.1401
    s = 0.8811 1.2516 1.116 0.7513

  sigma:  0.0384

     AIC     AICc      BIC
402.0749 410.2567 415.2665

        Ljung-Box test

data:  Residuals from Holt-Winters' multiplicative method
Q* = 45.657, df = 3, p-value = 6.708e-10

Model df: 8.   Total lags used: 11

Error measures:
                    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
2020 Q4       2665.314 2534.085 2796.544 2464.616 2866.013
2021 Q1       2303.393 2189.951 2416.836 2129.898 2476.889
2021 Q2       3467.136 3296.324 3637.948 3205.901 3728.370
2021 Q3       3939.817 3745.648 4133.986 3642.861 4236.773
2021 Q4       2809.789 2671.256 2948.321 2597.921 3021.656
```

Fig. 12. Summary of Holt Exponential for NewHouseReg Data

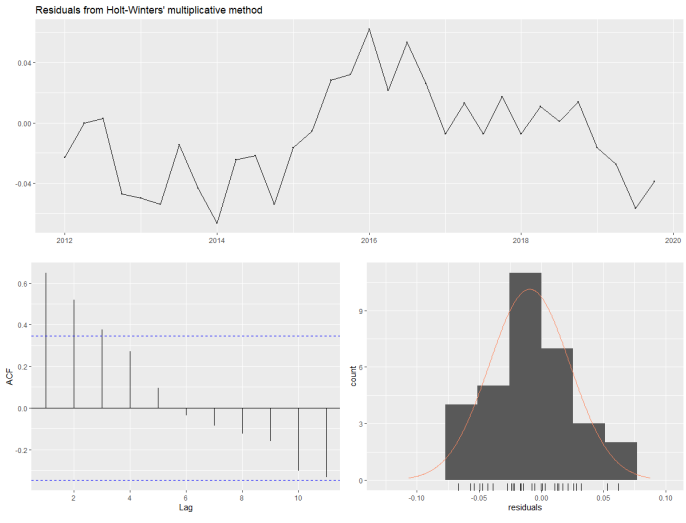


Fig. 13. Residual Plot of Holt Exponential for NewHouseReg Data

(ARIMA):Non-seasonal type of ARIMA is appropriate and used here because the data is in trending pattern and does not have any seasonality. To work with ARIMA, the data should be stationary and hence with using ndiffs() function got to know that the data is already stationary as ndiffs() returned 0. The values p,d,q were extracted manually, p value was extracted using Pacf(ts_NewHouseReg) and graph was plot, it was noticed that partial auto correlation factor as lag at 2 as seen in fig.15 where as in autocorrelation factor there is a decay in lag which forms a sine wave as seen in fig.14 hence it was observed that it is an auto regression ARIMA model so the value of p will b 2, d will be 0 as suggested by ndiffs() and q will be 0 according to the ACF chart. These values were passed to model to make it fit as this Arima(ts_NewHouseReg, c(2,0,0)) and a forecast plot was created as seen in the fig.16 In this dataset also the ARIMA model performed exponential good and was one of the best model compared to previous models as the RMSE score is 6342.208, p-value is 0.2344 and as per residual plot its been observed that all lags are inside the significant boundary all residuals are normally as well as the AIC score is also best.

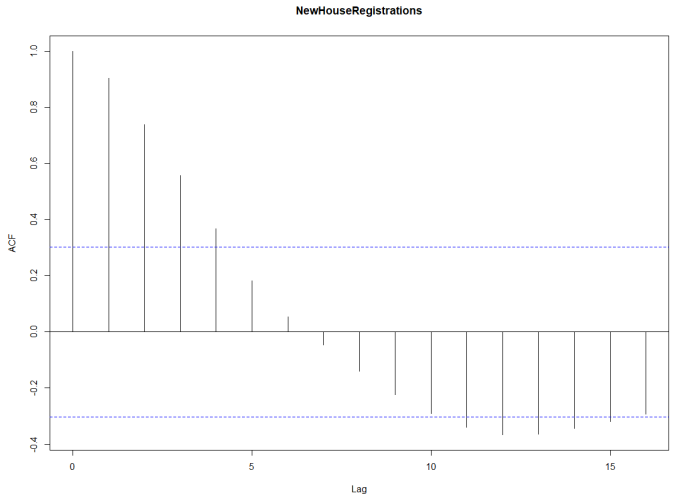


Fig. 14. ACF chart for NewHouseReg Data

B. Dataset 2: OverseasTrips

1) Data Description and Data Understanding: The OverseasTrips datasets consists of 32 records and 2 features also its been observed that trips events are carried out quarterly. In order to work with time series in r, the raw data needs to be converted to time series object in R using ts() function.

As the observations in data is occurring every 4 months thus frequency = 4 is given as well as the data is not numeric vector thus a start vector is been created as the data starts with 2012Q1 so start = c(2012,1) is given.

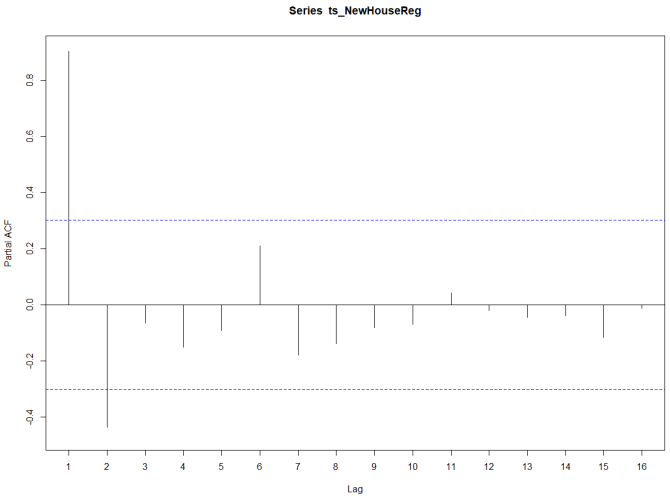


Fig. 15. PACF chart for NewHouseReg Data

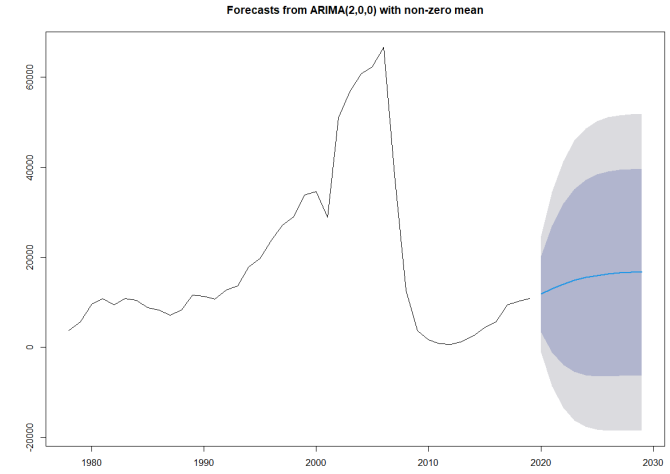


Fig. 16. Forecast Plot of ARIMA for NewHouseReg Data

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

Coefficients:
         ar1      ar2       mean
      1.3346  -0.4665  16791.106
s.e.  0.1315   0.1319   6985.181

sigma^2 estimated as 43317727:  log likelihood=-428.43
AIC=864.86   AICc=865.94   BIC=871.81

Training set error measures:
                    ME     RMSE      MAE       MPE     MAPE      MASE         ACF1
Training set 207.1252 6342.208 3464.418 -20.20197 35.95662 0.8732557 -0.007018081
> checkresiduals(arimamodpred)

        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
```

Fig. 17. Summary of ARIMA for NewHouseReg Data

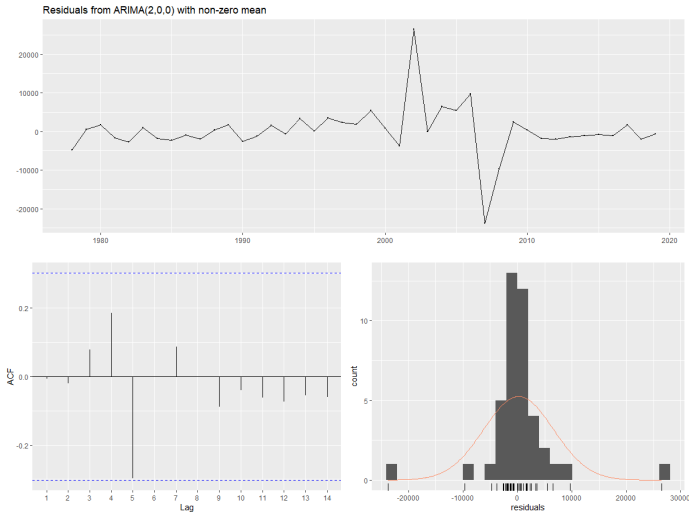


Fig. 18. Residual Plot of ARIMA for NewHouseReg Data

```
> ts_OverseasTrips <- ts(data2, start = c(2012,1), frequency = 4)
> ts_OverseasTrips
     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
```

Fig. 19. Time Series Data

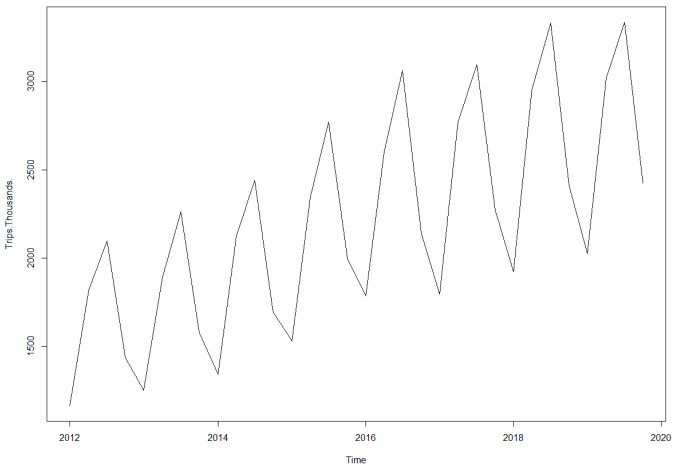


Fig. 20. OverseasTrips Data

After creating time series data, the data was than plotted as seen in fig.20 to check whether the data is time series trend data seasonal time series data or is the data stationary meaning that the observations are constant and it's been observed that the data is trending as a linear trend can be seen as well as it is seasonal data as seasonality is fixed and known frequency because the observations are dropped at some fixed quarters also some observations are peaked during some fixed quarters.

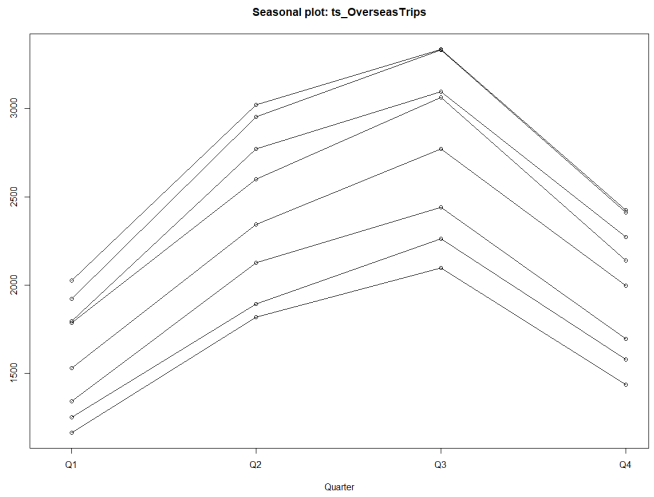


Fig. 21. Seaonalplot of OverseasTrips Data

A seasonal plot is plotted(fig.21) to check the trend of each individual season. A seasonal plot is somewhat like time plot. Seasonal plot helps to understand seasonal pattern more clearly. It's been observed that for every year there is an increase in number of trips till quarter 2 followed by more increase in number of trips in quarter 3 and during quarter 4 the number of trips to Ireland got decreasing.

Once got to know that the data is of combination of seasonal and trend pattern, the next step was to check whether the data is stationary or no. In order to check if data is stationary ndiffs() function was used in R and got to know that the data was not stationary as it returned 1. So the data was made stationary by differentiating and the trend was removed. After making the data stationary the data was plotted as seen in fig.22 and its been observed that the data is stationary and in multiplicative way as the amplitudes are in increasing fashion.

As the data is in multiplicative in nature seasonal decomposition is performed using multiplicative type method, decompose(data,type= "multiplicative") function is used in R and a plot is created as seen in fig.23 and it demonstrates that the data consists on a trend pattern as its increasing over period of time in a linear fashion, also the plot helps to clarifies that the data is in seasonal pattern as well as there is some randomness, these random/error/ Irregular are patterns

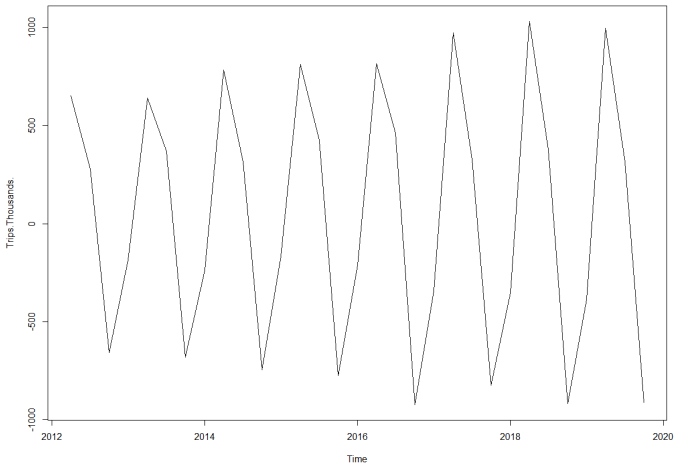


Fig. 22. Stationary data of OverseasTrips Data

that can't be explained.

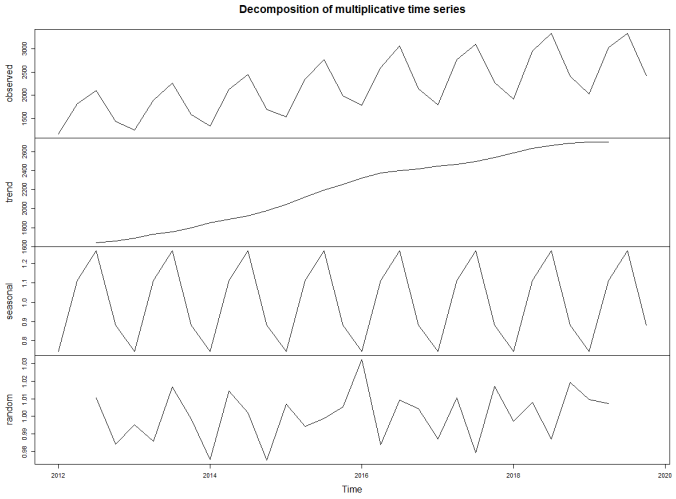


Fig. 23. Decomposition using Multiplicative model for OverseasTrips Data

2) Models:

- Simple Average mean: The data was build using meanf(data,h=3) model in R, here we want to forecast future 3 period so have passed h=3. The fig.24 demonstrates that the model has predicted future three quarters and their forecasted values are 2209 as seen in the fig.25 After plotting forecast plot residuals were checked using Ljung-Box test to verify whether that our model was good and acceptable or not, by seeing the summary of the average mean, it can be understood that the model was not good as we got RMSE 598 and idle it's good if RMSE score is least or minimum as well as the p-value should be greater than 0.05 but the model gave 1.168e-07. Thus will drop this model as it is not acceptable as well as residual plot seen in fig.26 helps to confirms the same as through ACF it can be seen that the lags are going out of the significant boundary line.

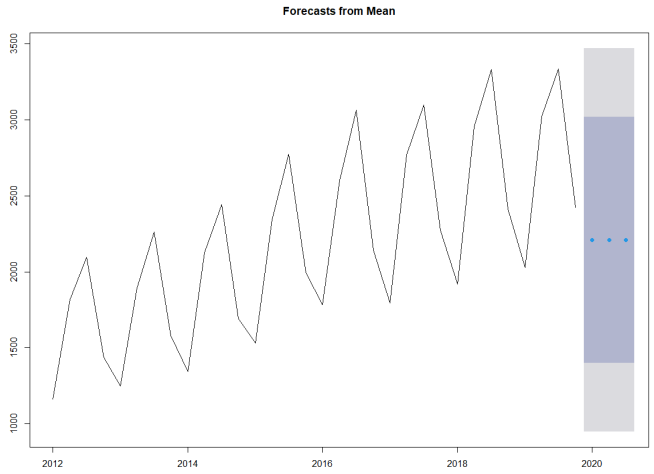


Fig. 24. Forecast Plot of Average Mean Model for OverseasTrips Data

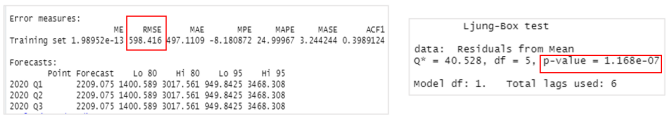


Fig. 25. Summary of Average Mean Model for OverseasTrips Data

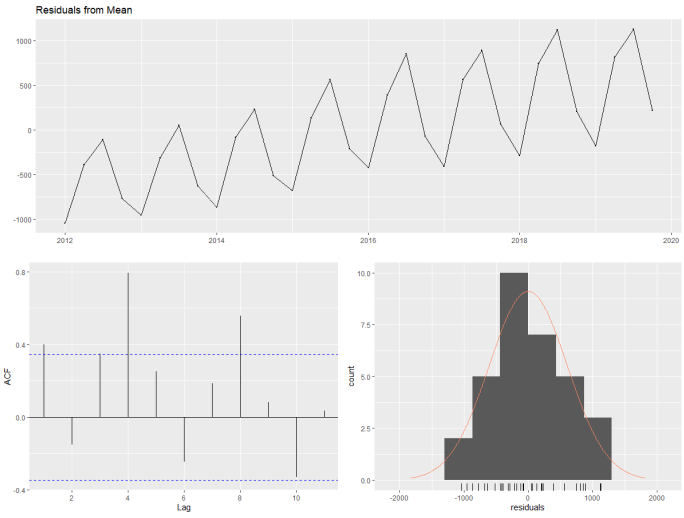


Fig. 26. Residual Plot of Average Mean Model for OverseasTrips Data

- Naïve seasonal: Naïve takes random walk it either goes randomly uptrend or down trend. snaive(data,h=3) is used to build model in R, h indicates number of periods to be predicted. The fig.27 Naïve seasonal forecast plot demonstrates that 3 random observations are ploted based on past pattern. The fig.27 demonstrates that 2026,3021,

3334 for Q1, Q2 and Q3 are future 3 forecasted values. The fig.28 is summary of naïve seasonal model and it's been observed that RMSE got decreased and now it's 176 as well as by Ljung-box test the p-value is improved and its 0.01 but still it's not greater than 0.05 so the model is not acceptable. Also as per residual plot in fig.29 it shows that as per ACF there is one lag going outsit significant boundary so this naïve seasonal is not the best model and will drop this model and look for another model.

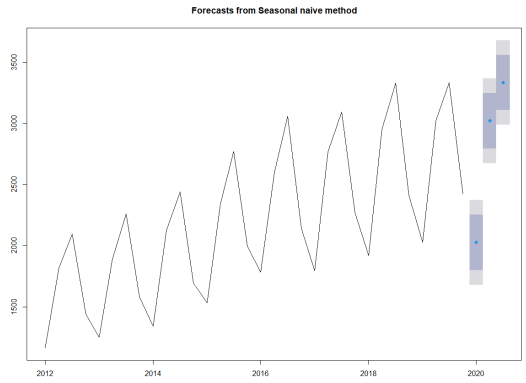


Fig. 27. Forecast Plot of Naive Seasonal Model for OverseasTrips Data

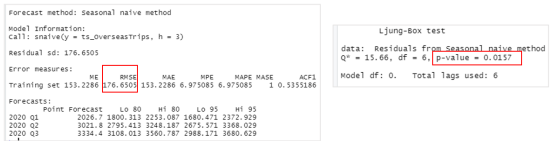


Fig. 28. Summary of Naive Seasonal Model for OverseasTrips Data

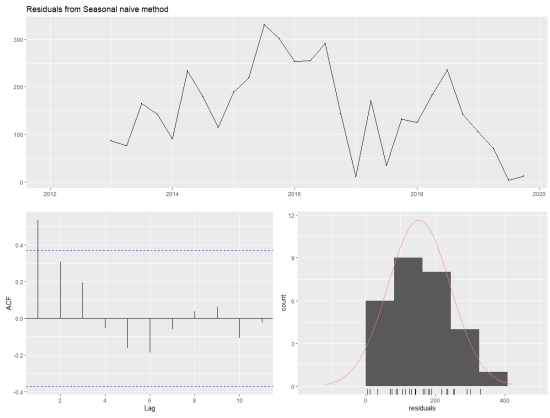


Fig. 29. Residual Plot of Naive Seasonal Model for OverseasTrips Data

- Holt's - winter exponential smoothing: As the data is seasonal and multiplicative in nature holt's winter model is used. Although this model performed somewhat similar to naïve seasonal model but RMSE i.e 72 as seen in fig.31 was improved also the p-value. A residual plot(fig.32) was plotted to check whether the lag in inside the significant boundary or not, and it was observed that there were some lags going outside significant boundary although residuals were normalized. This model was kept on hold and another model was build using ARIMA to check whether we can get the best model or no.

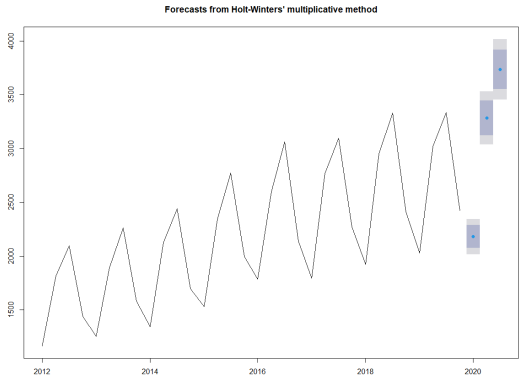


Fig. 30. Forecast Plot of Holt's - winter exponential Model for OverseasTrips Data

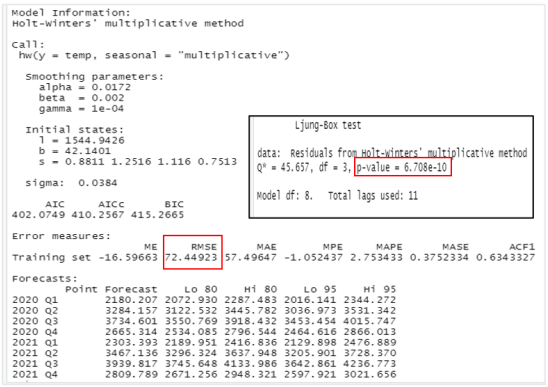


Fig. 31. Summary of Holt's - winter exponential Model for OverseasTrips Data

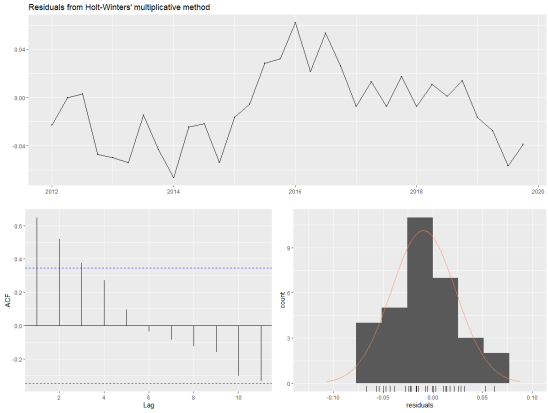


Fig. 32. Residual Plot of Holt's - winter exponential Model for OverseasTrips Data

- Seasonal ARIMA(SARIMA): SARIMA is appropriate to use when the data has seasonal pattern thus model was used. There are few steps in ARIMA modelling which were done such as the raw data was plotted first, than the data was checked for stationaty using ndiffs() function adf.tests(ts) and got to know that the data is not stationary, so the data was made stationary by using diff(ts, differences=d) , where d indicates the number of times the time series ts is differenced. Once the data was made stationary, autoARIMA was executed to check p,d,q and P,D,Q values and it suggested ARIMA(1,0,0)(0,1,0)[4] with drift. These values were than passed AS arimasfit¡-Arima(ts_OverseasTrips,order = c(1,0,0),seasonal = c(0,1,0)) to fit the model. ARIMA performed exponentially amazing compared to previous models. The fig.34 shows that RMSE score was least i.e 72 also p-value was greater than 0.05 which 0.16 which is very good and acceptable on the other hand it's observed that AIC is also very good.

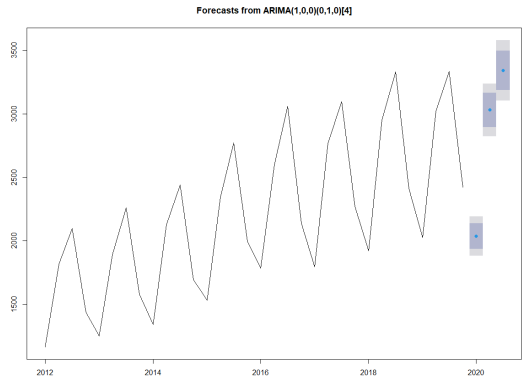


Fig. 33. Forecast Plot of SARIMA Model for OverseasTrips Data

```
> summary(arimasfit)
Series: ts_OverseasTrips
ARIMA(1,0,0)(0,1,0)[4]

Coefficients:
          ar1
      0.8762
s.e.  0.0755

sigma^2 estimated as 6256:  log likelihood=-162.33
AIC=328.66   AICc=329.14   BIC=331.32

Training set error measures:
                    ME     RMSE      MAE       MPE     MAPE      MASE      ACF1
Training set 15.71939 72.65164 54.33896 0.6292887 2.385652 0.3546268 -0.268398

        Ljung-Box test

data:  Residuals from ARIMA(1,0,0)(0,1,0)[4]
Q* = 7.9014, df = 5, p-value = 0.1618

Model df: 1.   Total lags used: 6
```

Fig. 34. Summary of SARIMA Model for OverseasTrips Data

3) Summary and Results of Part A: As per the table fig.36, we can conclude that for NewHouseRegistrations dataset the ARIMA model of type non-seasonal performed very well.Whereas, on the other hand for OverseasTrips dataset again the ARIMA model was the best model as shown in the fig.37

II. Part B: Logistic Regression

A. Dataset – ChildBirth Dataset

1) Data description and understanding: For Part B, Logistic regression was performed on childbirth dataset in in

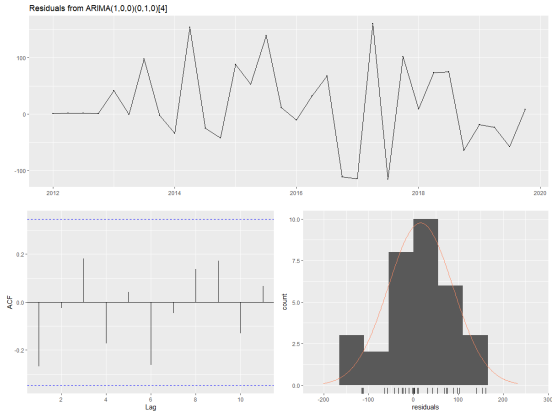


Fig. 35. Residual Plot of SARIMA Model for OverseasTrips Data

Methods	RMSE	MAE	MAPE	AIC
Mean Method	17881.98	14062.65	271.8443	-
Naïve (Random Walk)	7466.737	3967.244	34.07645	-
Simple Exponential	7378.822	3875.789	33.33161	911.1171
Holt's Exponential	7509.436	4276.987	87.12899	916.591
Non -Seasonal ARIMA	6342.208	3464.418	35.95662	864.86

Fig. 36. Results & comparission of NewHouseRegistrations mode

a US city. The dataset consists of 16 features and 42 records, where smoker, lowbwt and mage35 were 3 factor(categorical) variables and remaining were numeric.

For given dataset, decriptive statistics(fig.38) was created using SPSS, which demosnstrates mean value, Standard Error as well as Standard Deviation and Variance for all the variables.

Methods	RMSE	MAE	MAPE	AIC
Mean Method	598.416	497.1109	24.99967	-
Seasonal Naïve	176.6505	153.2286	6.975085	-
Holt's Winter	72.44923	57.49647	2.753433	402.0749
Seasonal ARIMA	72.65164	54.33896	2.385652	328.66

Fig. 37. Results & comparission of OverseasTrips model

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean		Std. Deviation	Variance
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic
ID	42	1737	27	1764	894.07	72.155	467.616	218664.897
Length	42	15	43	58	51.33	.453	2.936	8.618
Birthweight	42	2.65	1.92	4.57	3.3129	.09318	.60390	.365
Headcirc	42	9	30	39	34.60	.370	2.400	5.759
Gestation	42	12	33	45	39.19	.408	2.643	6.987
smoker	42	1	0	1	.52	.078	.505	.256
mage	42	23	18	41	25.55	.874	5.666	32.107
mnocig	42	50	0	50	9.43	1.931	12.512	156.544
mheight	42	32	149	181	164.45	1.004	6.504	42.303
mppwt	42	33	45	78	57.50	1.111	7.198	51.817
fage	42	27	19	46	28.90	1.059	6.864	47.113
fedyrs	42	6	10	16	13.67	.333	2.160	4.667
fnocig	42	50	0	50	17.19	2.671	17.308	299.573
fheight	42	31	169	200	180.50	1.077	6.978	48.695
lowbwt	42	1	0	1	.14	.055	.354	.125
mage35	42	1	0	1	.10	.046	.297	.088
Valid N (listwise)	42							

Fig. 38. Decriptive Statistics for ChildBirth data

For the given dataset a histogram was plotted as seen in fig.39, to check all variables and summarize discrete or continuous data as well as seeing the main data points. Thus, its been observed that the lowbwt and mage35 are imbalced wheres smoker is balaced.

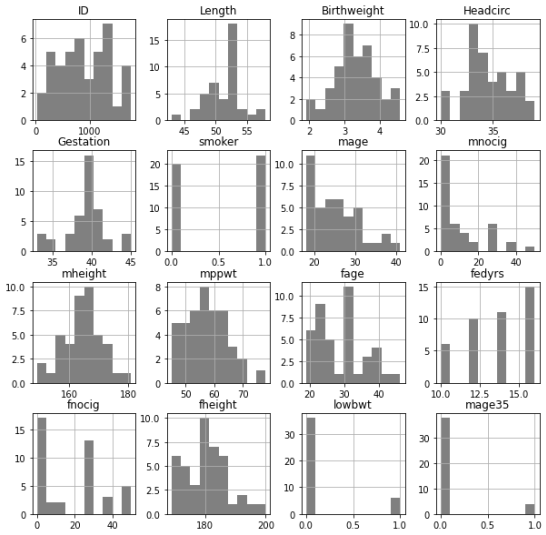


Fig. 39. Histogram for ChildBirth data

Before applying model it is necessary to check if the dataset consists of any outlier, hence a box plot was plotted to check if there are any outliers and it was observed that there were some outliers in length, gestation, mnocing, mppwt, fage, lowbwt and mage35. Although, these outlier were not harmfull for building model as well as removing outliers was not good option as the records were less so outlier were not removed.

This dataset does not consists of a target column and to perform logistic regression a target variable is necessary. So by understanding data and analysing data's behavior the lowbwt was chose as target varibale to classify whether if a baby weight is low or not,where 0 means child birth weight was not less and 1 means child births weight was low. Rational for choosing lowbwt was because there are few independent columns which were supporting to lowbwt such as smoker column, where 1 indicates mother is smoker and 0 indicates that mother is not smoker and mage35 column which dsispalys mothers age over 35, where 0 means NO and 1 means YES.

2) Feature selection: After choosing the target variable the next step was to do feature selection, a logistic model cannot be build using all variables, if done the model would be bias and won't predict the right outcome because there must some variables which are of no use and are not helping for prediction. Thus, feature selection was done using Principal Component Analysis(PCA), PCA helps to give the best features that are actually helping and contributing in building the best model. But, is PCA really advised or no for this dataset was checked with KMO & Barlett's Test in SPSS and it was suggested that PCA can be done, the values of test are shown in the fig.40 which is 0.552 and which is actually good because the idle value should be greater than 0.5 After applying PCA,

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.552
Bartlett's Test of Sphericity	Approx. Chi-Square	249.411
	df	78
	Sig.	.000

Fig. 40. KMO & Barlett's Test

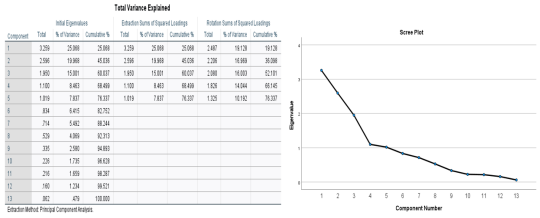


Fig. 41. Total variance and scree plot

5 variables were best for building the model which are shown in the left side of the fig.41 also looking at the spree plot shown in the right side of the fig.41 it can be observed that Eigenvalues are dipped upto 4 components, after 4 component there is deviation so 3 components were considered and PCA was recalculated using the 3 components.
A correlation matrix for all the variable was created after again doing PCA and extracting 3 components. This helped to understand whether any variable was getting correlated to each other.

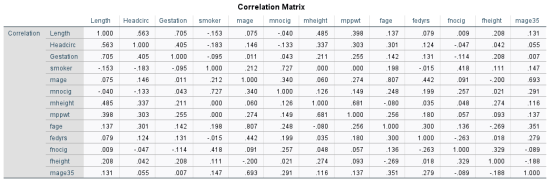


Fig. 42. correlation matrix

Finally, as shown in fig.43 three components were generated using Varimax rotation and Kaiser normalization, and these components were then fed into a Logistic Regression model.

3) Model Building, Evaluation and Results:

- Model 1: At this stage the best 3 features were extracted and fed as input to Logistic Regression Model and we got the result as shown in the fig.44 as classification table. The model performed very well and gave accuracy of 97.6. The model predicted 1 false negative, the model had to predict that the child birth weight is low but the model predicted that the baby is not low weight. It was observed that sensitivity was 83.3, means rate of postive

Component Transformation Matrix

Component	1	2	3
1	.681	.698	.223
2	-.707	.546	.450
3	.193	-.464	.865

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

Fig. 43. Component Tranforamtion Matrix

predictions which are correct where as on the other hand specificity was 100%, means rate at which true positive are correct.
Variable in the equation fig.45 helps to understand that the second component has high significance so the 2nd component was dropped and the model was build again.

Classification Table[a]

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

a. The cut value is .500

Fig. 44. Classification Table for model 1

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	REGR factor score 1 for analysis 2	-4.659	2.371	3.861	1	.049	.009
	REGR factor score 2 for analysis 2	-.856	1.018	.708	1	.400	.425
	REGR factor score 3 for analysis 2	2.718	1.617	2.826	1	.093	15.149
	Constant	-6.101	2.717	5.041	1	.025	.002

a. Variable(s) entered on step 1: REGR factor score 1 for analysis 2, REGR factor score 2 for analysis 2, REGR factor score 3 for analysis 2.

Fig. 45. Variable in the equation model 1

- Model 2: After dropping 2nd component, there were only two components 1st and 3rd to fed the model and the model gave the results shown in fig.46 as classification table and it was observed that even after dropping the second component the model was performing same and it did not help in predicting the model , it again gave false negative value.
  Variables in the equation fig.47 helps again to know that are both component significance was less than 0.05 and hence this was the best model and acceptable

Classification Table[a]

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

a. The cut value is .500

Fig. 46. Variable in the equation for model 2

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	REGR factor score 1 for analysis 2	-4.020	1.614	6.203	1	.013	.018
	REGR factor score 3 for analysis 2	2.467	1.237	3.974	1	.046	11.782
	Constant	-5.620	2.130	6.960	1	.008	.004

a. Variable(s) entered on step 1: REGR factor score 1 for analysis 2, REGR factor score 3 for analysis 2.

Fig. 47. Variable in the equation for model 2

B. Summary

Thus, it can be concluded that for NewHouseRegistrations datasets, non-seasonal ARIMA model performed very well and was best model among other models as well as for OverseasTrips dataset the model was performing good with Seasonal ARIMA. Whereas, for logistic regression, the model was helpful to predict with 2 components which had significance value less than 0.05 and it was acceptable and best model.

REFERENCES

[1] vshkapil, "Optimized Solution," Wordpress.com. [Online]. Available: https://optimizerkapil.wordpress.com/).