Executive Summary

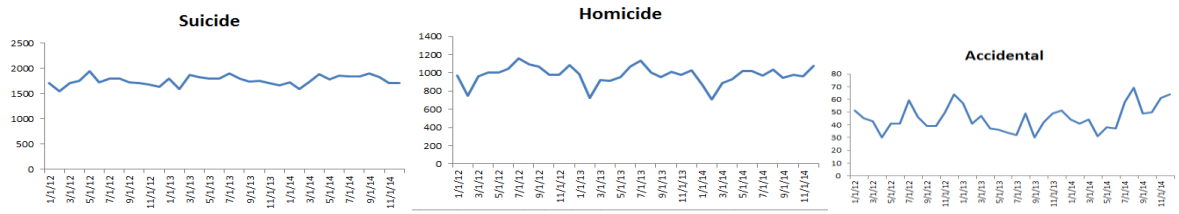
This report focusses on creating monthly forecasts of suicides using firearms for the year 2015. Action Alliance is a big organization with operations in various social areas. With such large requirements of contractual work force, there is **huge scope of cost savings** by efficient human resource allocation.

This forecasting exercise predicts with 95% accuracy, monthly suicides involving firearms. The model also predicts the deaths by gender, age and location of death. With a year's view of the deaths statistics, Action alliance will be able to point out months of low and high deaths thereby pre-preparing for contingencies. Efficient allocation also improves effectiveness of counselling as the right volunteers can be procured by observing categorized series (gender, age, location of death).

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Suicide - Gender - Male	+2.81%	+1.94%	+4.62%	-2.01%	+4.46%	-0.97%	+2.90%	+1.42%	-3.88%	-1.59%	+0.97%	-0.91%
Suicide - Gender - Female	+7.95%	-1.81%	+1.53%	+2.74%	+7.26%	-2.66%	-1.91%	-7.53%	-7.21%	-3.09%	+2.72%	+5.75%
Suicide - Age - 0 to 20	+3.32%	+14.22%	+7.44%	+2.88%	+22.02%	+4.68%	+24.50%	-1.12%	+10.69%	+0.96%	-0.08%	+21.58%
Suicide - Age - 21 to 40	+4.69%	+4.22%	+5.66%	-1.42%	+5.43%	+2.84%	+5.59%	-0.36%	+1.67%	+0.93%	-1.37%	-5.00%
Suicide - Age - 41 to 60	+2.52%	-2.73%	+1.78%	-0.99%	-2.11%	-5.42%	-1.05%	-0.97%	-5.29%	-3.28%	+5.53%	-3.27%
Suicide - Age - 61 to 100	+7.91%	+7.16%	+9.88%	+1.89%	+15.62%	+4.44%	+6.09%	+7.35%	-4.25%	+3.20%	+6.38%	+13.30%
Suicide - Place - Farm	+16.97%	-16.33%	+96.32%	+2.30%	-7.92%	+4.81%	-11.86%	+35.82%	-28.48%	-6.48%	-31.77%	-15.66%
Suicide - Place - Residence	+3.38%	+1.70%	+6.85%	+1.16%	+6.70%	+2.28%	+7.21%	+5.02%	-0.52%	+2.39%	+6.41%	+5.05%
Suicide - Place - Institution	-14.85%	-17.98%	+18.95%	+24.73%	+11.34%	-10.24%	+7.58%	-12.24%	+13.24%	-14.51%	+16.23%	-1.61%
Suicide - Place - Workplace	+11.08%	+11.94%	+1.66%	-2.77%	+7.60%	-3.12%	-4.22%	-5.57%	-6.24%	-2.86%	-1.77%	-2.57%
Suicide - Total	+5.21%	+3.43%	+6.26%	+0.66%	+7.19%	+1.20%	+4.78%	+2.84%	-1.58%	+1.24%	+4.65%	+3.53%

The above table shows the %increase in number of suicides compared to the same month in the last year. There is a marked increase in the months of February and May for the total number of suicides. Similarly, there is a marked increase of 14% in the number of suicides committed by youngsters under the age of 20. There is a 22% forecasted increase in the number of suicides in May within the same age group. Similar trends can be observed by location of the incident.

Data Source and Description – Overall firearm deaths data (monthly data from 2012 to 2014) was sourced from Kaggle (https://www.kaggle.com/hakabuk/gun-deaths-in-the-us) for this analysis containing over 100K deaths. Data was filtered for suicides (up to 66% cases are suicides, rest being homicides and accidents). There are an average of **1800 monthly suicides**, **~1000 homicide incidents** and **~50 accidents** involving firearms. Forecasts predict an average increase of 4% in suicides in the next year. Charts below show the overall gun deaths split by suicides, homicides and accidents. Graphs show yearly seasonality, no trend and noise in few months. It has details of each individual who died in this period – age of the person, race, place of death, educational qualification and gender. It comprised of suicide, homicide and accidental cases of deaths by firearms.



Forecasting Methods - Forecasts were generated using naïve, smoothing and multiple linear regression models. Final forecasts were created after comparing a host of methods (details below) and choosing data points from most favorable methods. Various models were trained using data for 24 months and then validated using another 12 months. Same models were then used to create another 12 months of forecasts. Upon comparing all methods, Holt's Winter Additive method was found to be the most suitable candidate for forecasting.

Conclusion and Recommendation

Monthly suicide numbers should be used for planning and setting up contracts for sourcing of contract employees an year ahead of time saving considerable costs. Unusual busy periods could be identified for better readiness. Though the model forecasts for 12 months into the future, this exercise should be repeated every 6 months to include actual data and re-create future forecasts for planning considering the fact that the socio-economic and political space in undergoing rapid change in the US at present.

Technical Summary

This section describes the technical details of the project – the data, preparation steps, forecasting methods and model evaluation.

Data Preparation – The data was clean and didn't require any cleaning steps. As part of data preparation, the suicide data was filtered out and bucketed by demographics of the deceased – age group, gender and place of death on a monthly basis. The groups used under each category are mentioned below:

- Gender: Male, Female
- Age: 0 – 20, 21 – 40, 41 – 60, 61 – 100
- Place of death: Residence, Workplace, Institutions, Farms (Farms category was not merged with workplace as farms represented rural areas while the workplace data was that of urban areas).

Hence, after this step each category had 36 data points (3 years monthly data). **Appendix 1** shows the plot of these data. None of the months had any missing values.

Description of Prepared Data – From the graphs in **Appendix 1** of the prepared data, a weak increasing trend and annual seasonality can be inferred. The seasonality is more pronounced in

some, such as count of deceased males, and almost absent in some cases, such as count of deceased females.

Forecasting and Modeling – Based upon the trend and seasonality in data, different forecasting approaches were tried out. Forecasting was done for each category and the total count. Hence, in total 11 series were forecasted – 2 for gender, 4 for age, 4 for place of death, and 1 for total number of suicides. As the goal is to forecast number of suicide attempts for the next 1 year in each category, the data set was divided into training data set of 2 years (24 data points) and into validation data set of 1 year (12 data points) for each. The different forecasting methods used have been described in detail below:

1. **Naïve (Seasonal)**: Different naïve approaches were used to forecast the series. However, the one that worked the best was with a seasonality of 12 months. This model serves as the **benchmark** for other models. This model performed the best in case of forecasting male deaths (**Table 1**). **Appendix 3** has relevant plots.
2. **Multiple Linear Regressions**: MLR with combinations of linear trend, quadratic trend and seasonality were used for each series. This method gave the best MAPE for forecasting suicides in the age group of 0 to 20. The details of the model are given below. As the RMSE of training and validation data sets is comparable, hence there is no overfitting. **Appendix 3** has relevant plots.

Regression Model

Input Variables	Coefficient	Std. Error	t-Statistic	P-Value	CI Lower	CI Upper	RSS Reduction
Intercept	60	10.23270899	5.86354992	0.000109	37.47796	82.52204	125860.2
t	0.111111	0.372160028	0.298557348	0.770842	-0.70801	0.93023	20.09043
Jan	22.72222	11.68013442	1.945373349	0.077728	-2.98558	48.43002	248.2577
Feb	7.611111	11.55495458	0.658688103	0.523644	-17.8212	33.04339	20.81285
Mar	13.5	11.44051645	1.180016659	0.262883	-11.6804	38.68041	15.43213
Apr	28.38889	11.33714533	2.504059714	0.029294	3.436	53.34178	758.1653
May	13.27778	11.24514642	1.180756327	0.262601	-11.4726	38.02818	81.7617
Jun	-0.33333	11.16480084	-0.02985573	0.976717	-24.9069	24.24023	109.1315
Jul	-3.44444	11.09636174	-0.31041205	0.76205	-27.8674	20.97848	307.1301
Aug	3.944444	11.04005055	0.357284999	0.727639	-20.3545	28.24343	99.08133
Sep	21.33333	10.99605359	1.940089976	0.07843	-2.86882	45.53548	244.9641
Oct	14.22222	10.9645191	1.297113179	0.221138	-9.91052	38.35497	99.35875
Nov	11.11111	10.94555479	1.015125439	0.331854	-12.9799	35.20211	123.3141

Residual DF	11
R^2	0.617771
Adjusted R^2	0.200794
Std. Error Estimate	10.93923
RSS	1316.333

Training Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
1316.333	7.405891	-3.55271E-15

Validation Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
913.5	8.724964	3

3. **Holt's Winter Additive**: This method performed the best for age groups 41 to 60, 61 to 100, place of death – institution and the total number of suicides (from **Table 1**). The additive approach performed better than the multiplicative one; this is also intuitive from the data plots where the seasonality appears to be more of additive type than of multiplicative type. As mentioned in the executive summary section, this is also the model that is proposed for the client. **Appendix 2** gives the details of HW additive models for all the series. From the data, it is observed that the MAPE scores of training and validation sets are comparable and hence, there is almost no indication of over-fitting. The constants – alpha, beta and gamma were evaluated by a trial and error approach for each series.
4. **Double Exponential**: Since some of the series such as deaths at workplace and farms had a trend with no clear seasonality, double exponential method was also used as a forecasting approach. The constants – alpha and beta were evaluated by a trial and error approach for each series. The method performed the best for forecasting suicide attempts in the age group 21 to 40 and at farm

locations (from **Table 1**). The details of these two models are given below. **Appendix 3** has relevant plots.

Inputs

Data	
Workbook	SH - Location.xlsx
Worksheet	Data_PartitionTS
Range	$B$20:$BF$56
Selected Variable	S_Place_Farm_fix
# Records in Training Data	24
# Records in Validation Data	12

Parameters/Options	
Optimization Selected	No
Alpha (Level)	0.2
Beta (Trend)	0.15
Forecast	Yes
#Forecasts	12

Training Error Measures

Mean Absolute Percentage Error (MAPE)	34.31716095
Mean Absolute Deviation (MAD)	3.2480878
Mean Square Error (MSE)	17.7648341
Tracking Signal Error (TSE)	2.536382126
Cumulative Forecast Error (CFE)	8.238391837
Mean Forecast Error (MFE)	0.343266327

Validation Error Measures

Mean Absolute Percentage Error (MAPE)	19.62301398
Mean Absolute Deviation (MAD)	2.512246194
Mean Square Error (MSE)	10.77483203
Tracking Signal Error (TSE)	7.801233841
Cumulative Forecast Error (CFE)	19.59862002
Mean Forecast Error (MFE)	1.633218335

Inputs

Data	
Workbook	SH - Age.xlsx
Worksheet	Data_PartitionTS
Range	$B$20:$BS$56
Selected Variable	S_Age_21-40_fix
# Records in Training Data	24
# Records in Validation Data	12

Parameters/Options	
Optimization Selected	No
Alpha (Level)	0
Beta (Trend)	0.17
Forecast	Yes
#Forecasts	12

Training Error Measures

Mean Absolute Percentage Error (MAPE)	4.19062061
Mean Absolute Deviation (MAD)	19.45089358
Mean Square Error (MSE)	598.5587403
Tracking Signal Error (TSE)	4.164831319
Cumulative Forecast Error (CFE)	81.00969076
Mean Forecast Error (MFE)	3.375403782

Validation Error Measures

Mean Absolute Percentage Error (MAPE)	3.831904168
Mean Absolute Deviation (MAD)	17.58008829
Mean Square Error (MSE)	545.5888155
Tracking Signal Error (TSE)	-3.235934412
Cumulative Forecast Error (CFE)	-56.88801265
Mean Forecast Error (MFE)	-4.740667721

5. **Simple Exponential**: Series of female suicides, suicides at institutions follow have almost no trend and seasonality. Hence, this method was used to primarily forecast such series; but other series were also forecasted for the possibility to come up with a model that has least MAPE when compared to models of other methods. The level constant – alpha was determined using a hit and trial approach; starting off with a seed value suggested by optimize option. This model was the best to forecast suicides at workplace (from **Table 1**). The details of the model are given below for this series. Since the MAPE for validation is lower than that of training, there is no over-fitting in this model. **Appendix 3** has relevant plots.

Inputs

Data	
Workbook	Suicide - Exp.xlsx
Worksheet	Data_PartitionTS
Range	$B$20:$L$56
Selected Variable	S_Place_WorkPlace
# Records in Training Data	24
# Records in Validation Data	12

Parameters/Options	
Optimization Selected	No
Alpha (Level)	0.34
Forecast	Yes
#Forecasts	12

Training Error Measures

Mean Absolute Percentage Error (MAPE)	6.669518473
Mean Absolute Deviation (MAD)	28.65225428
Mean Square Error (MSE)	1518.664611
Tracking Signal Error (TSE)	-5.121709145
Cumulative Forecast Error (CFE)	-146.7485127
Mean Forecast Error (MFE)	-6.38037012

Validation Error Measures

Mean Absolute Percentage Error (MAPE)	5.627608693
Mean Absolute Deviation (MAD)	19.84056965
Mean Square Error (MSE)	853.5625837
Tracking Signal Error (TSE)	-4.512018363
Cumulative Forecast Error (CFE)	-89.52101457
Mean Forecast Error (MFE)	-7.460084548

6. **Moving Average**: The absence of visible trend and seasonality in series of female suicides, suicides at institutions was the motivation to try this method. Different window sizes were tried, but the window size of 12 gave the best results. This method did not give the best results for any series in particular; however, it was the closest to the best for forecasting suicides at far location. Below are the details of the models used for the farm series. Since the MAPE for training and validation data sets are similar, it is likely that there is no over-fitting.

Inputs

Data	
Workbook	Suicide.xlsx
Worksheet	Data_PartitionTS
Range	$B$20:$L$56
Selected Variable	S_Place_Farm
# Records in Training Data	24
# Records in Validation Data	12

Parameters/Options	
Interval	12
Forecast	Yes
#Forecasts	12

Training Error Measures

Mean Absolute Percentage Error (MAPE)	22.78966643
Mean Absolute Deviation (MAD)	3.677778134
Mean Square Error (MSE)	9.454432179
Tracking Signal Error (TSE)	2.220713073
Cumulative Forecast Error (CFE)	9.083333333
Mean Forecast Error (MFE)	0.756944444

Validation Error Measures

Mean Absolute Percentage Error (MAPE)	20.56116723
Mean Absolute Deviation (MAD)	2.416666667
Mean Square Error (MSE)	8.138888889
Tracking Signal Error (TSE)	3.724137931
Cumulative Forecast Error (CFE)	9
Mean Forecast Error (MFE)	0.75

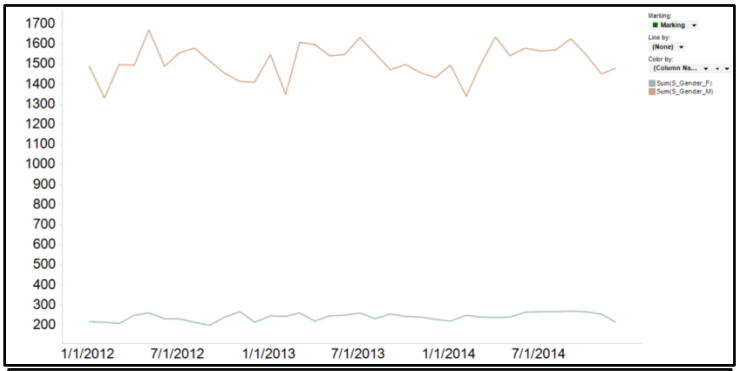
7. **Ensemble**: An ensemble approach was tried using all the 5 modelling methods and with 3 modelling methods (MLR, HW Additive, Double Exponential) separately. A uniform and weighted averaging approach was used. In case of weighted averaging approach, the weight of a method was calculated by:
   Weight = (Sum of absolute value of residuals for the method/ Sum of absolute value of residuals for all methods) x normalizing factor.
   The normalizing factor brought the weight in the range [0, 1] and ensured that sum of all weights is equal to 1. This method gave the highest weight to the method which generated least sum of errors across all the series.
   However, it was observed that the different types of ensembles didn't yield better results on the overall as shown in **Table 1**.
8. **Weighted forecasting approach**: In this approach the total number of suicide attempts was forecasted on a monthly basis and each category like age was calculated using a monthly weighted average of the past trend. For example, if the weight for male and female series (calculated from Jan 2012 and 2013 numbers) for Jan month are 0.8 and 0.2 respectively and the total number of suicides for Jan 2014 was forecasted to be 100, then the number of male and female suicide attempts in Jan 2014 were forecasted to be 80 and 20 respectively. However, this approach didn't give the best performance for any series as shown in **Table 1**.
   Note on auto-correlation, second layer models, and neural networks: Since the residuals obtained from any model didn't exhibit auto-correlation, a second layer model was not implemented in the project. As the size of the data is quite small (36 data points in total for each series), hence neural network was not used for modelling. Also, simpler methods such as HW additive have given fairly good results, so the need of a complicated model for the client was not realized. A simpler model such as HW additive has the added benefit of easy upkeep and low cost maintenance in the long run.
   Performance Evaluation – Performance evaluation of the models was done using MAPE and spread of residuals. As per **Table 1** below, HW additive method works best for the most number of series in terms of MAPE. This has been proposed to the client as a solution as mentioned earlier. **Appendix 2** shows the histogram of residuals and shows that models are more of over-estimating type; this is more favorable than under-estimating models as extra counsellors are better than vacancies for counselling to reduce the number of suicide attempts. **Appendix 4** shows the final forecasts for each of the series from the models.

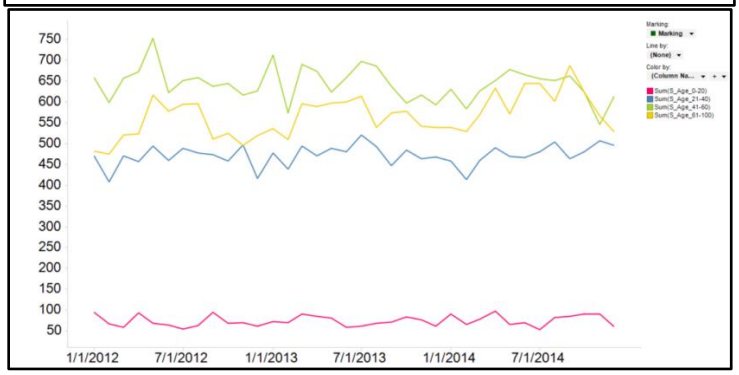
Annexures

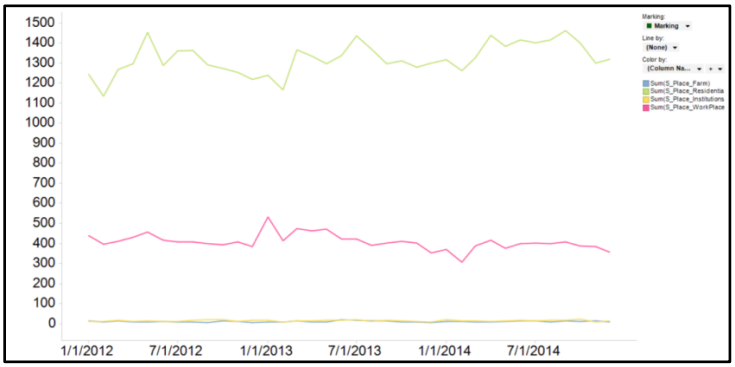
Table 1 – Comparison of Forecasting Approaches

	Naïve	MLR	Holt's Winter	Double Exponential	Simple Exponential	Moving Average	Ensemble	Weighted	Min MAPE
Suicide - Gender - Male	3.22%	3.59%	3.28%	4.71%	4.59%	4.29%	3.41%	3.18%	3.22%
Suicide - Gender - Female	6.61%	7.02%	6.53%	6.44%	6.84%	6.63%	6.19%	6.85%	6.19%
Suicide - Age - 0 to 20	14.24%	9.57%	9.78%	16.62%	16.62%	16.62%	12.14%	10.18%	9.57%
Suicide - Age - 21 to 40	4.85%	4.91%	4.22%	3.83%	4.04%	4.09%	4.30%	3.77%	3.83%
Suicide - Age - 41 to 60	6.13%	3.92%	3.73%	5.79%	4.87%	4.77%	3.81%	5.24%	3.73%
Suicide - Age - 61 to 100	6.08%	5.29%	5.04%	7.19%	9.25%	7.28%	5.54%	7.11%	5.04%
Suicide - Place - Farm	35.14%	22.49%	22.89%	19.62%	26.21%	20.56%	19.85%	25.41%	19.62%
Suicide - Place - Residence	5.17%	3.59%	3.32%	4.25%	5.36%	4.90%	2.75%	4.08%	2.75%
Suicide - Place - Institution	27.04%	19.70%	18.70%	24.23%	20.19%	20.53%	19.16%	18.65%	18.70%
Suicide - Place - Workplace	13.63%	10.91%	9.43%	7.00%	5.63%	13.05%	6.92%	12.36%	5.63%
Suicide - Total	3.40%	3.51%	3.22%	4.56%	4.50%	4.33%	3.59%		3.22%

Appendix 1 – Gender, age and place of death wise distribution of actual data



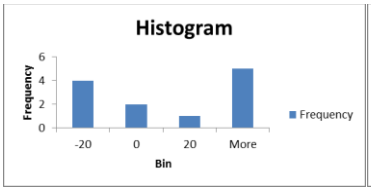




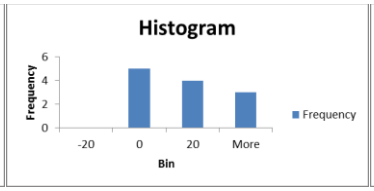
Appendix 2 – Holt's Winter Additive Models

	Gender		Age				Place of Death				Total
Training Error Measures	**Male**	**Female**	**0 - 20**	**21 - 40**	**41 - 60**	**61 - 100**	**Farm**	**Residence**	**Institution**	**Workplace**	
Mean Absolute Percentage Error (MAPE)	2.025786	5.837496	7.20985	2.589285	3.325877	3.046008	25.82936	1.831398	20.70365	3.822544	1.58504
Mean Absolute Deviation (MAD)	31.29076	13.87084	9.018721	12.00781	21.82537	17.04815	2.423018	24.28311	2.614552	16.35657	28.48445
Mean Square Error (MSE)	1458.142	244.9621	125.2586	204.6295	804.7292	448.9309	9.037725	1270.234	8.232231	475.7521	1761.937
Tracking Signal Error (TSE)	1.570519	1.644684	-5.24165	2.319551	-0.32954	-0.01589	0.100029	1.183446	-1.70324	-1.11107	0.099713
Cumulative Forecast Error (CFE)	49.14273	22.81316	-47.273	27.85273	-7.19234	-0.27084	0.242372	28.73775	-4.45321	-18.1733	2.840258
Mean Forecast Error (MFE)	2.047614	0.950548	-1.96971	1.16053	-0.29968	-0.01128	0.010099	1.197406	-0.18555	-0.75722	0.118344
Validation Error Measures											
Mean Absolute Percentage Error (MAPE)	3.277143	6.527454	8.492173	4.223664	3.7344	5.041397	22.89316	3.319671	18.70124	9.434768	3.217201
Mean Absolute Deviation (MAD)	51.02336	16.62785	10.74461	20.07201	23.10758	30.58593	2.41581	45.94503	3.004599	34.0124	57.79435
Mean Square Error (MSE)	3735.362	408.3246	218.7206	552.0231	842.1698	1872.881	10.28072	3016.807	13.73234	2573.412	4940.336
Tracking Signal Error (TSE)	2.726208	5.900206	2.775228	-1.55636	1.20542	1.024542	0.778389	5.681153	7.292278	-10.6948	-1.25647
Cumulative Forecast Error (CFE)	139.1003	98.10777	29.81874	-31.2392	27.85434	31.33658	1.88044	261.0208	21.91037	-363.757	-72.6169
Mean Forecast Error (MFE)	11.59169	8.175647	2.484895	-2.60327	2.321195	2.611382	0.156703	21.75173	1.825864	-30.3131	-6.05141
Parameters/Options											
Optimize Weights	No	No	No	No	No	No	No	No	No	No	No
Alpha (Level)	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.01
Beta (Trend)	0	0	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.22
Gamma (Seasonality)	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
Season length	12	12	12	12	12	12	12	12	12	12	12
Number of seasons	2	2	2	2	2	2	2	2	2	2	2
Forecast	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
#Forecasts	12	12	12	12	12	12	12	12	12	12	12

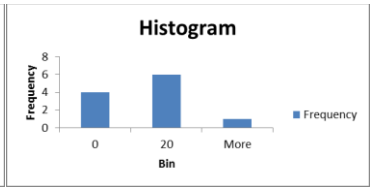
Residual plots



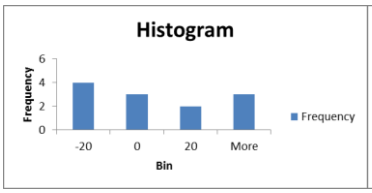
Gender – Male



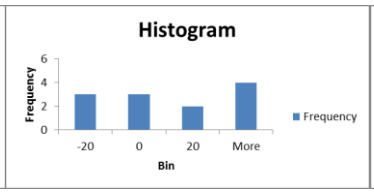
Gender – Female



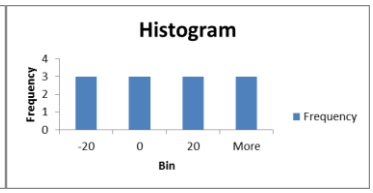
Age – 0 to 20



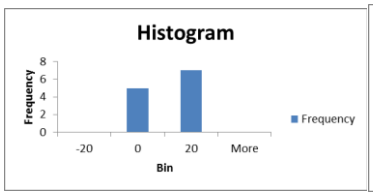
Age – 21 to 40



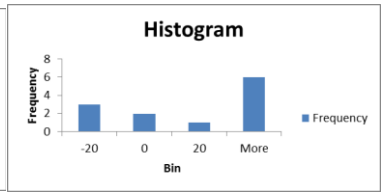
Age – 41 to 60



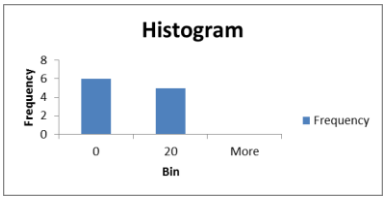
Age – 61 to 100



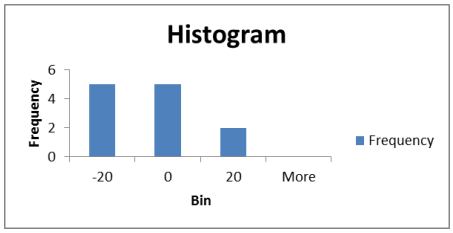
Place - Farm



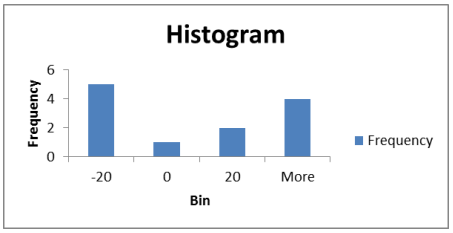
Place – Residence



Place – Institution

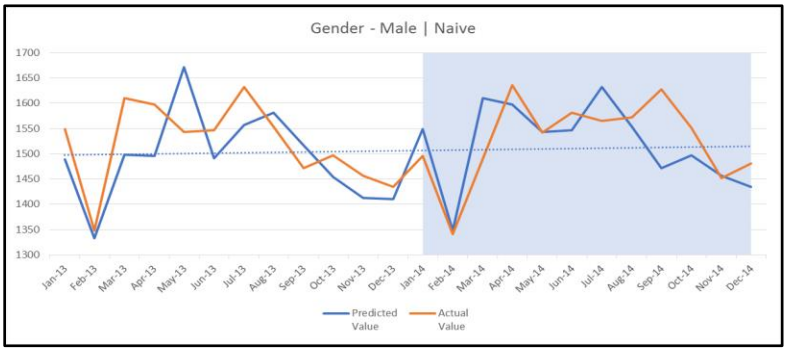


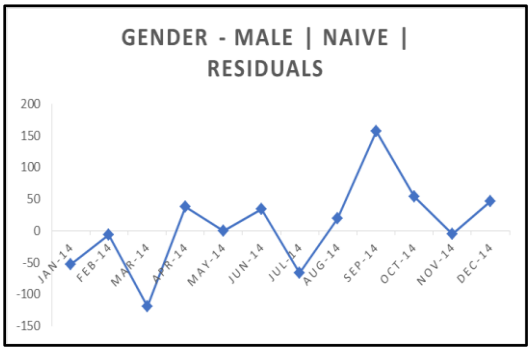
Place – Workplace

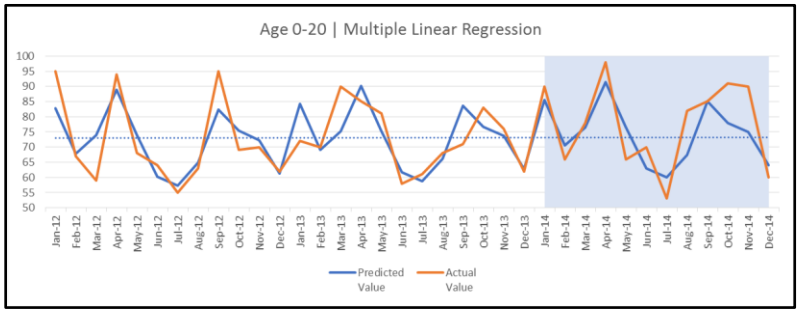


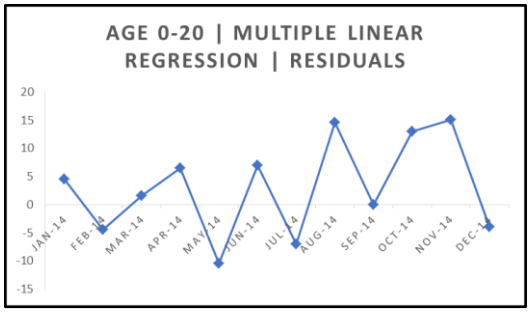
Total Suicide

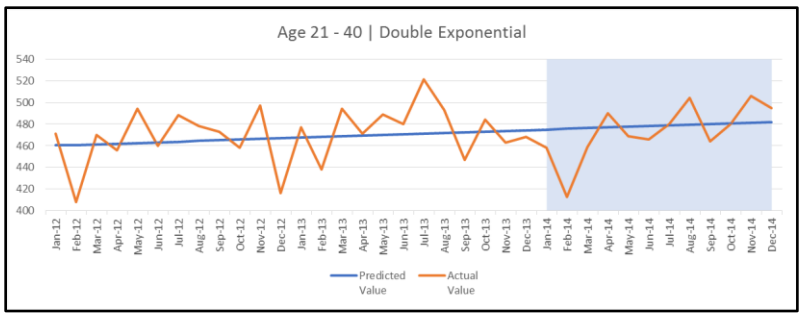
Appendix 3

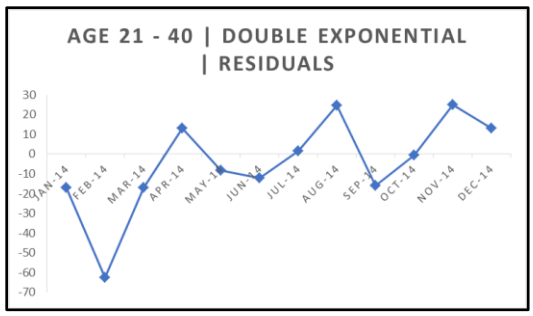

Gender - Male | Naive
Predicted Value
Actual Value

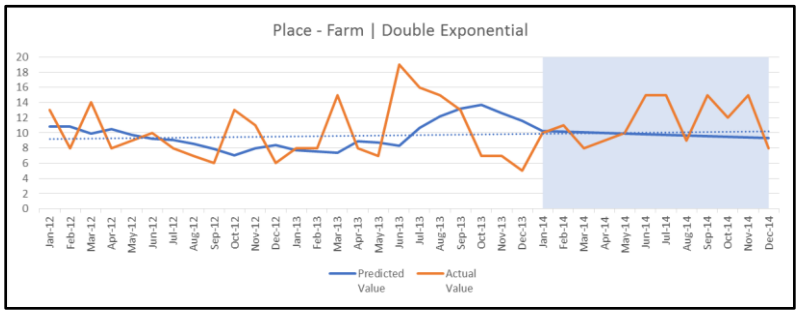

GENDER - MALE | NAIVE | RESIDUALS

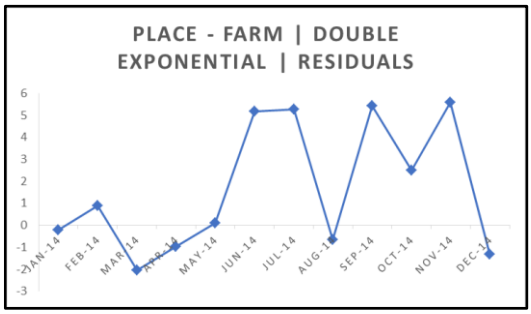

Age 0-20 | Multiple Linear Regression
Predicted Value
Actual Value

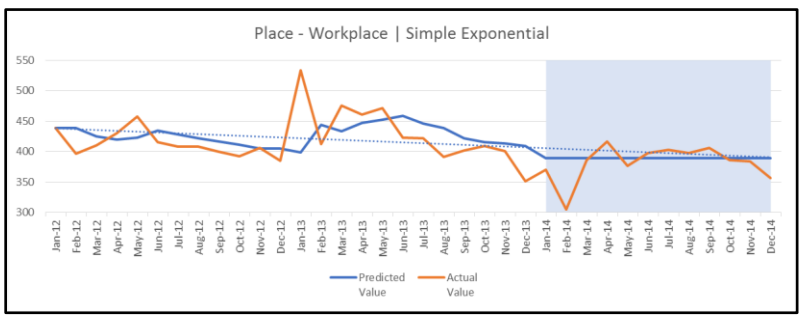

AGE 0-20 | MULTIPLE LINEAR REGRESSION | RESIDUALS

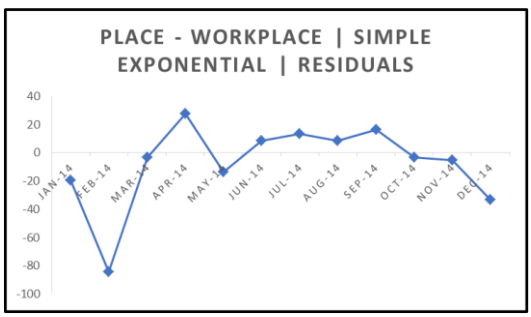

Age 21 - 40 | Double Exponential
Predicted Value
Actual Value


AGE 21 - 40 | DOUBLE EXPONENTIAL | RESIDUALS


Place - Farm | Double Exponential
Predicted Value
Actual Value


PLACE - FARM | DOUBLE EXPONENTIAL | RESIDUALS


Place - Workplace | Simple Exponential
Predicted Value
Actual Value


PLACE - WORKPLACE | SIMPLE EXPONENTIAL | RESIDUALS

Appendix 4 – Forecasts (Using Holt's Winter Additive)