US DEATH RECORDS ANALYSIS

OVERVIEW

- Critical to identify health issues
- Mortality rate - important indicator of the nation's health
- Various ongoing researches to identify & address primary cause of death
- Data source – Mortality dataset from CDC.gov (Centers for Disease Control and Prevention)
- 300,000 records
- 20 attributes

DATA CLEANING

- Step 1 - Merging of 1 main dataset with various lookup datasets

restatus	education	month_of_death	sex	age
1	3	1	M	1084
1	6	1	M	1070
1	3	1	F	1091
1	3	1	F	1040
1	5	1	F	1089
1	6	1	M	1078

Main Dataset

Resident Status Lookup Table

Code	Description
1	RESIDENTS
2	INTRASTATE NONRESIDENTS
3	INTERSTATE NONRESIDENTS
4	FOREIGN RESIDENTS

Education Lookup Table

Code	Description
1	8th grade or less
2	9 - 12th grade, no diploma
3	high school graduate or GED completed
4	some college credit, but no degree
5	Associate degree
6	Bachelor's degree
7	Master's degree
8	Doctorate or professional degree
9	Unknown

DATA CLEANING

- Step 2 – Converting Age formats into uniform format (i.e. Years)

age
1084
1070
1091
1040
1089
1078

Before conversion

Age	Age_type
84	Years
87	Years
90	Years
70	Years
58	Years
80	Years
91	Years

After conversion

1	001-135,999	...	Years
2	001-011,999	...	Months
4	001-027,999	...	Days
5	001-023,999	...	Hours
6	001-059,999	...	Minutes
9	999	...	Age not stated

Age Lookup Table

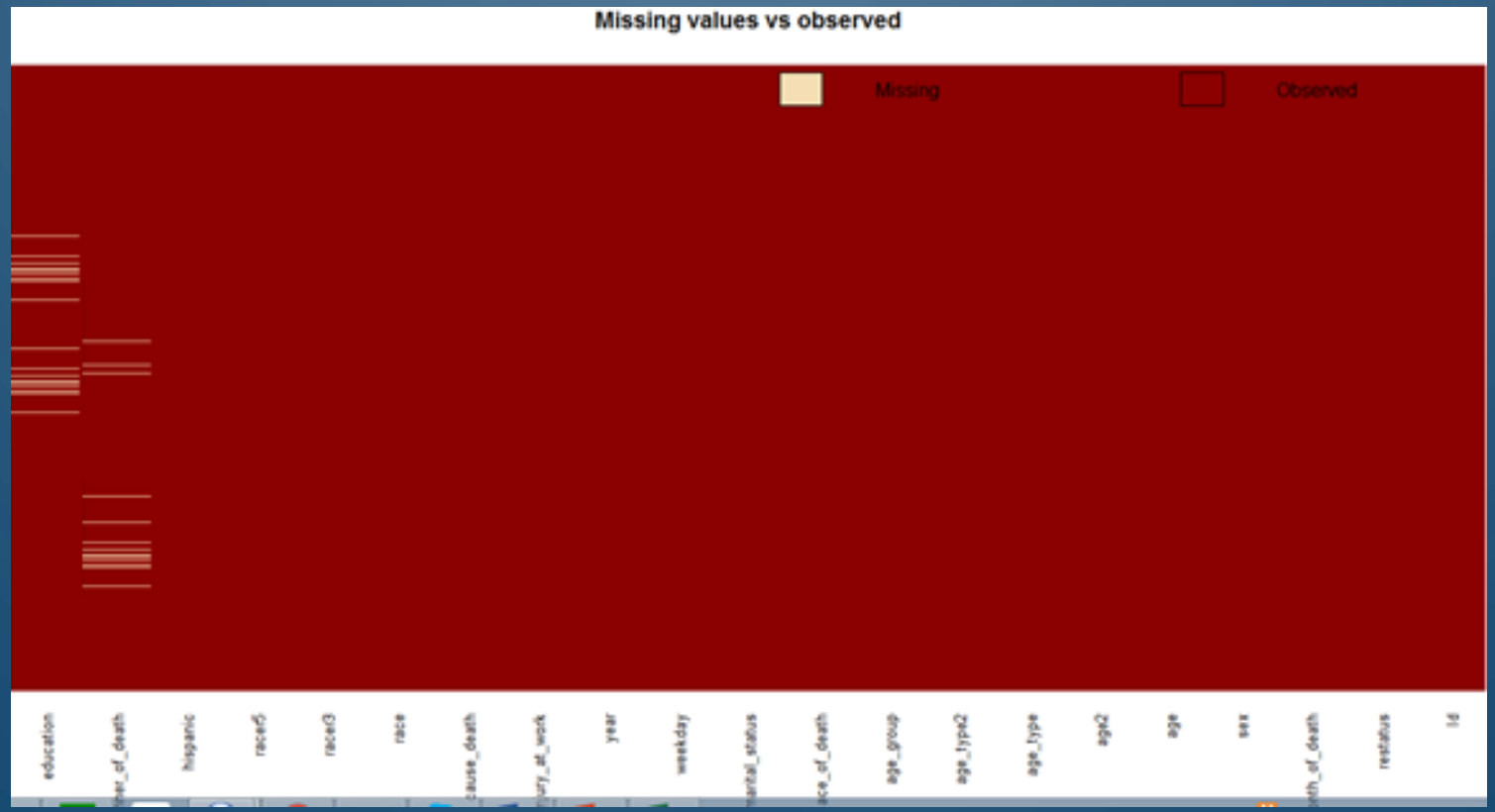
DATA CLEANING

- Step 3 – Grouping Age levels

```
data_all$AgeGroup <- cut(data_all$Age, c(0,19,25,39,60,110))
levels(data_all$AgeGroup) <- c("Teenager", "Young_Adult", "Adult", "Middle_Age","Senior_Citizens")
```

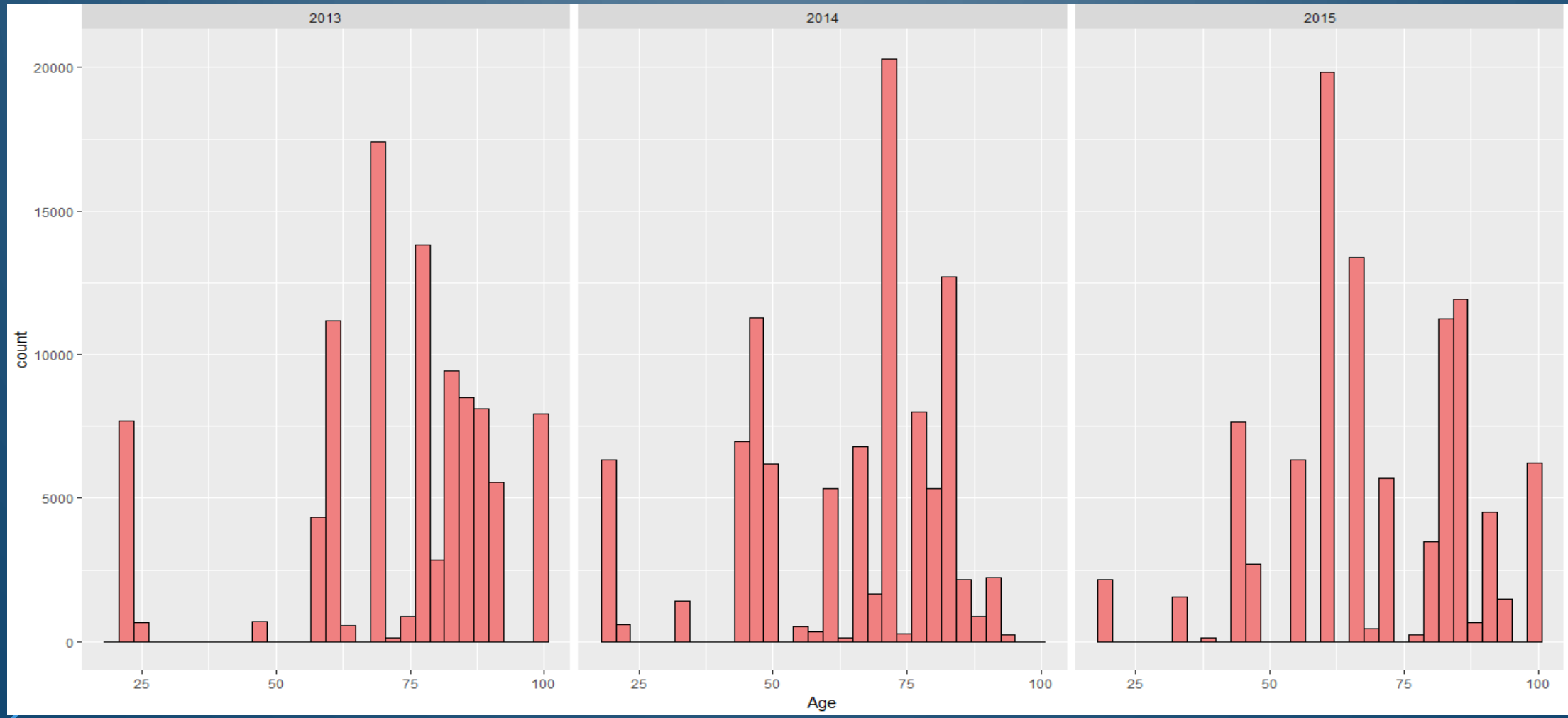
DATA CLEANING

- Step 4 – Check for Missing Values using “missmap”



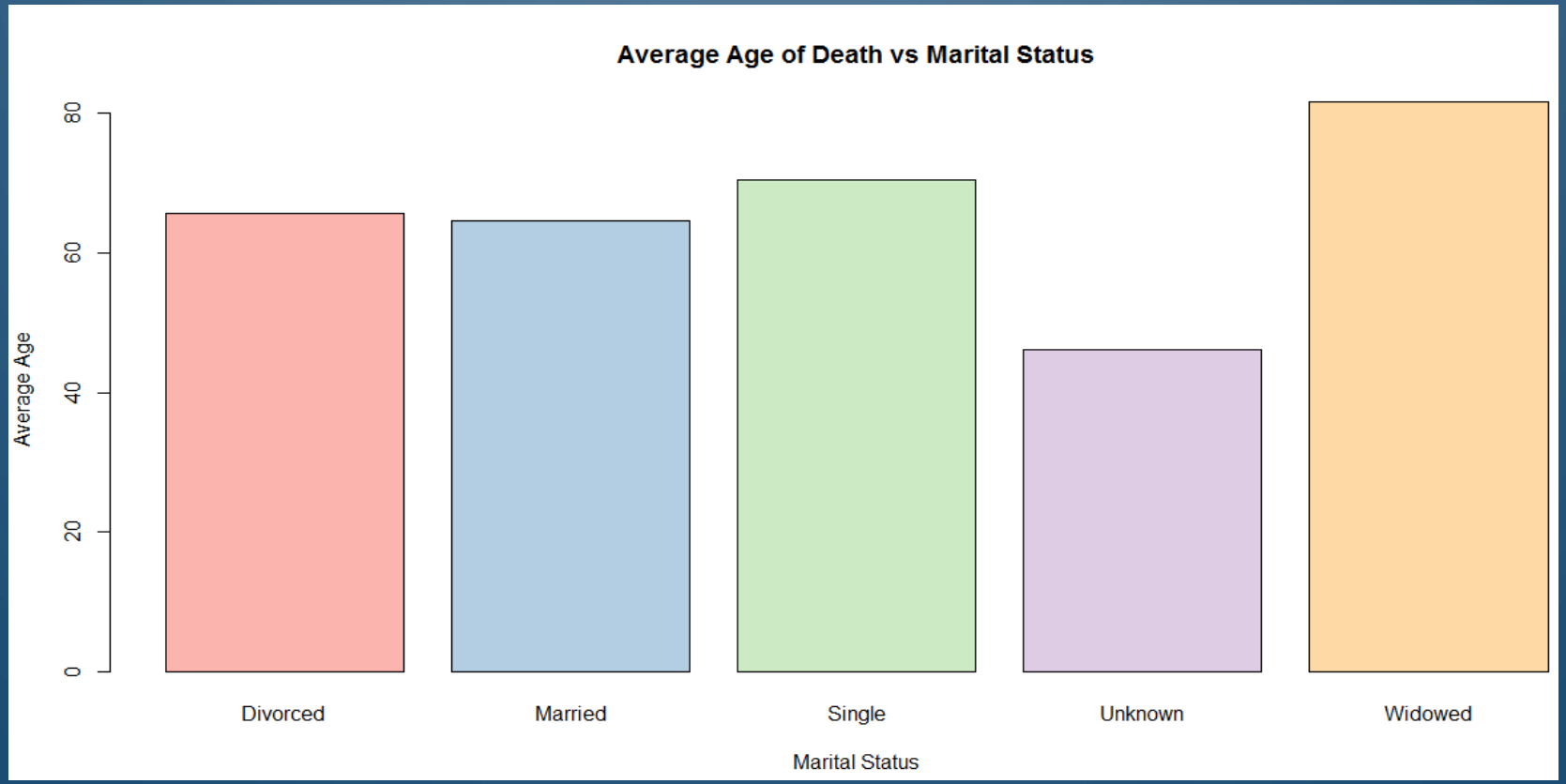
STATISTICAL ANALYSIS

- Death count according to Age and Years



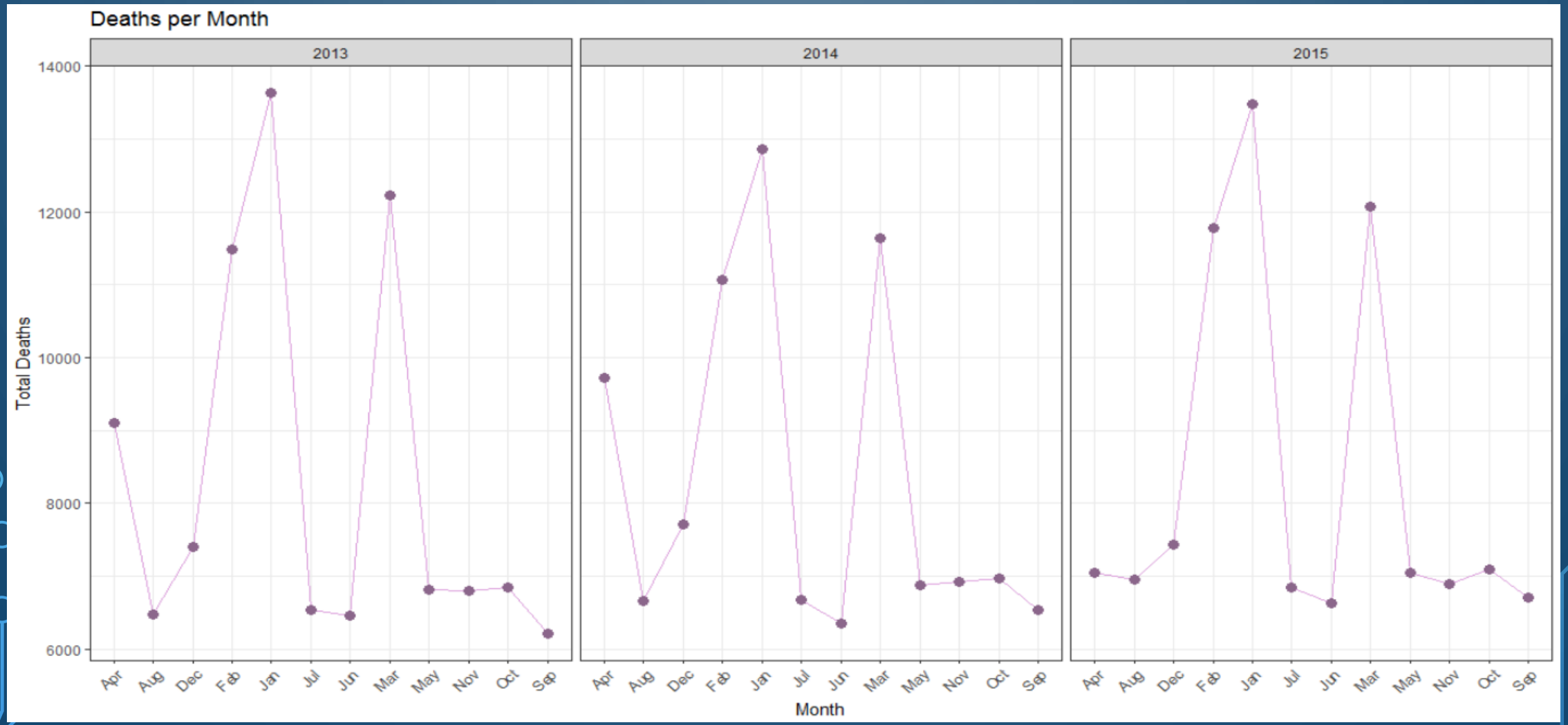
STATISTICAL ANALYSIS

- Average Age of Death according to Marital Status



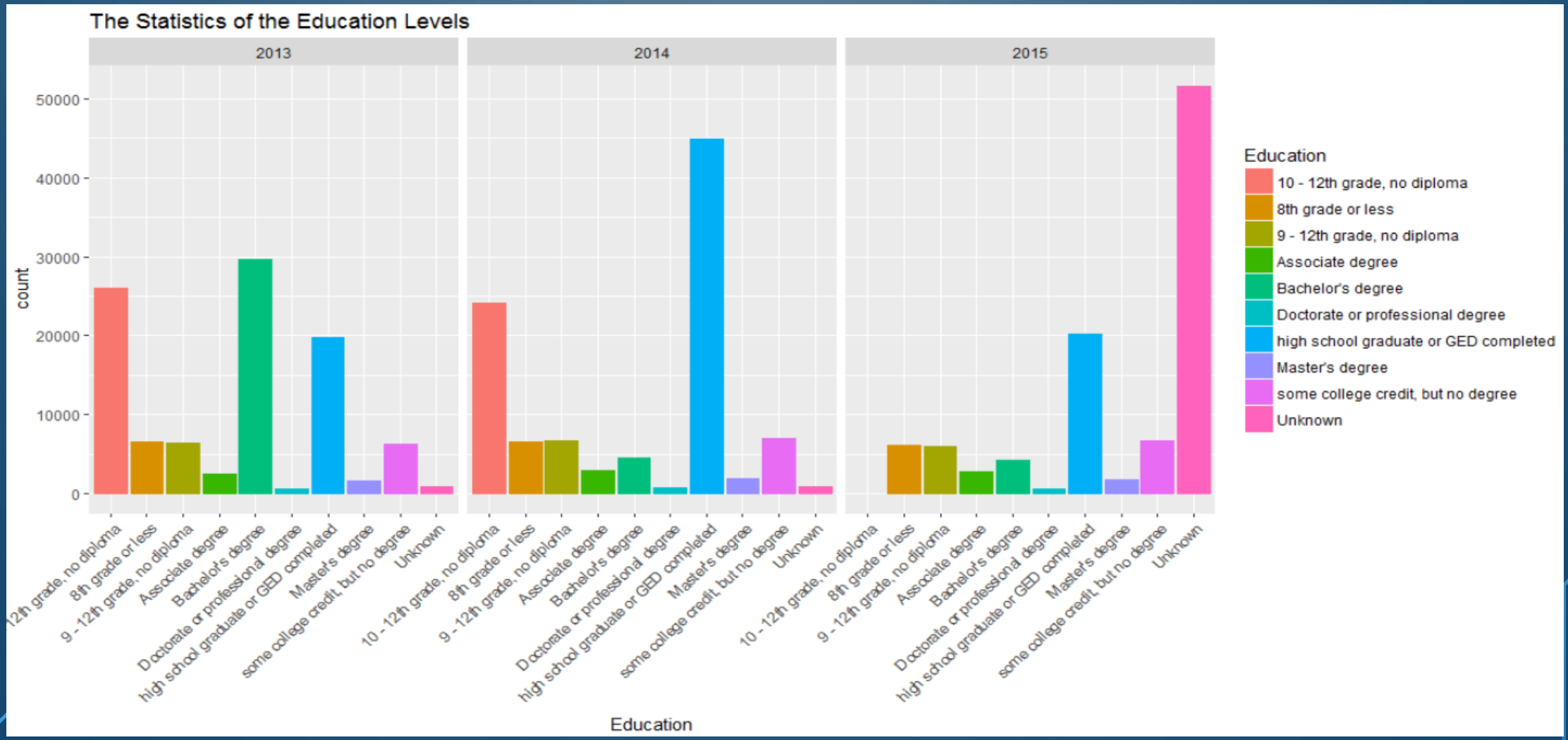
STATISTICAL ANALYSIS

- Pattern of Death per month throughout Years



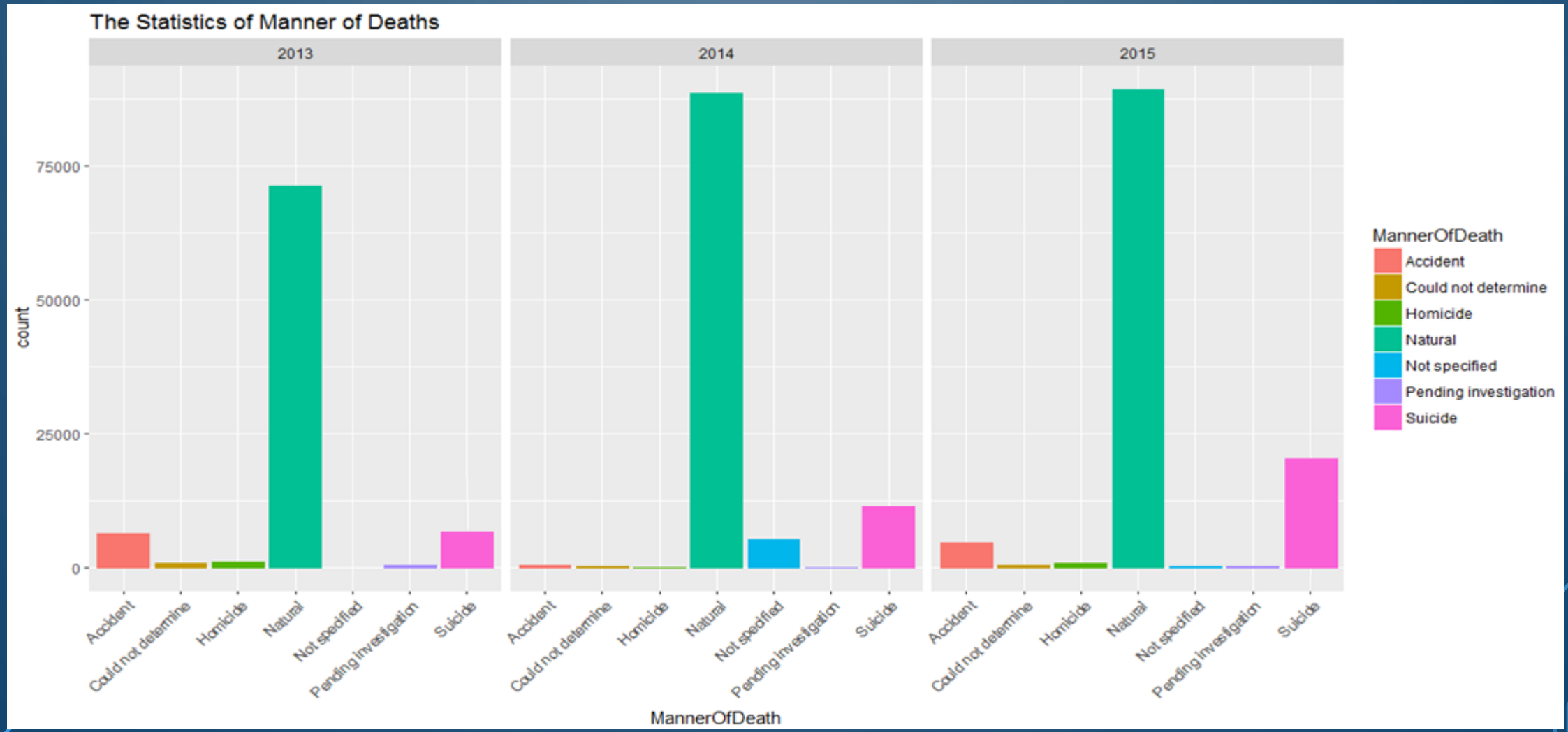
STATISTICAL ANALYSIS

- Death counts as per Education levels and Years



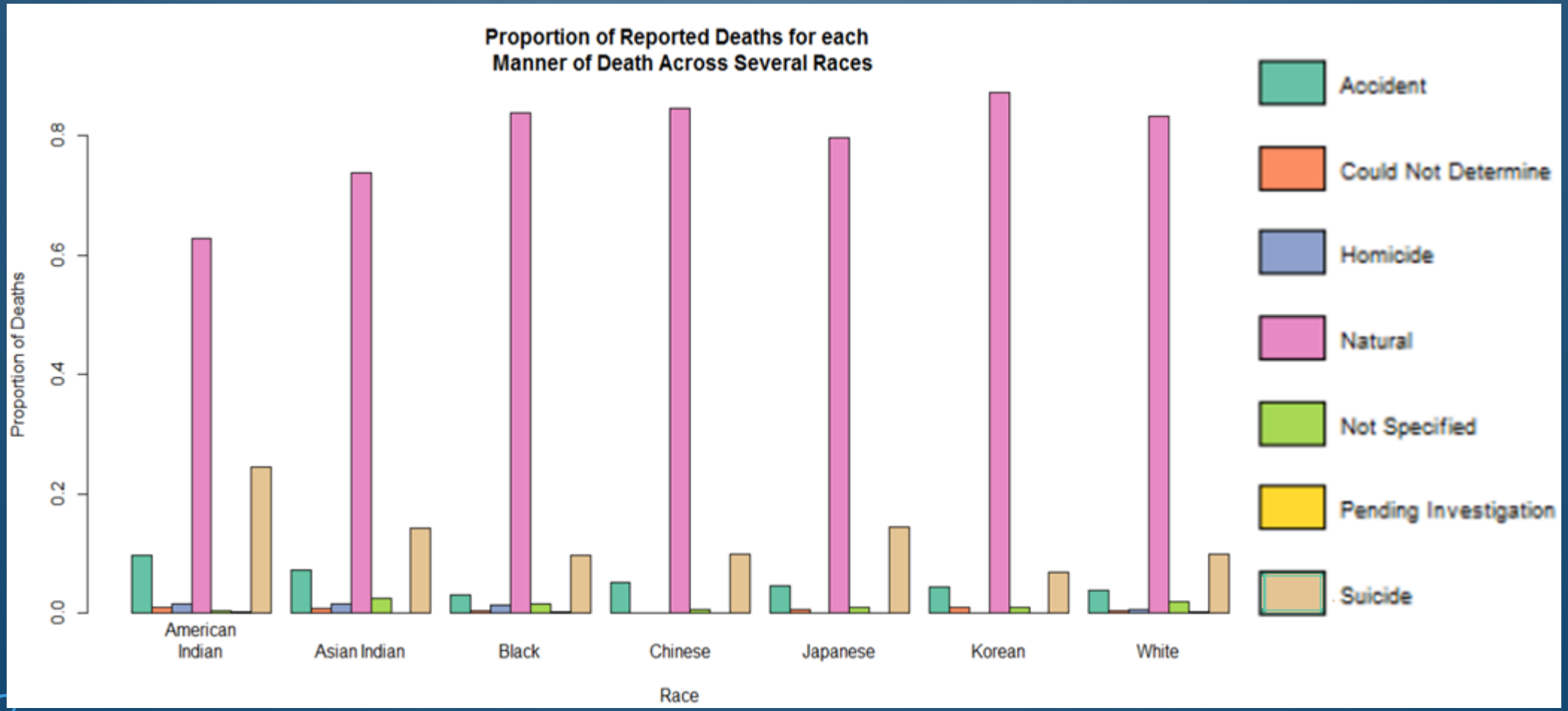
STATISTICAL ANALYSIS

- Pattern of Manner of Death according to Years – “Natural” & “Suicide”



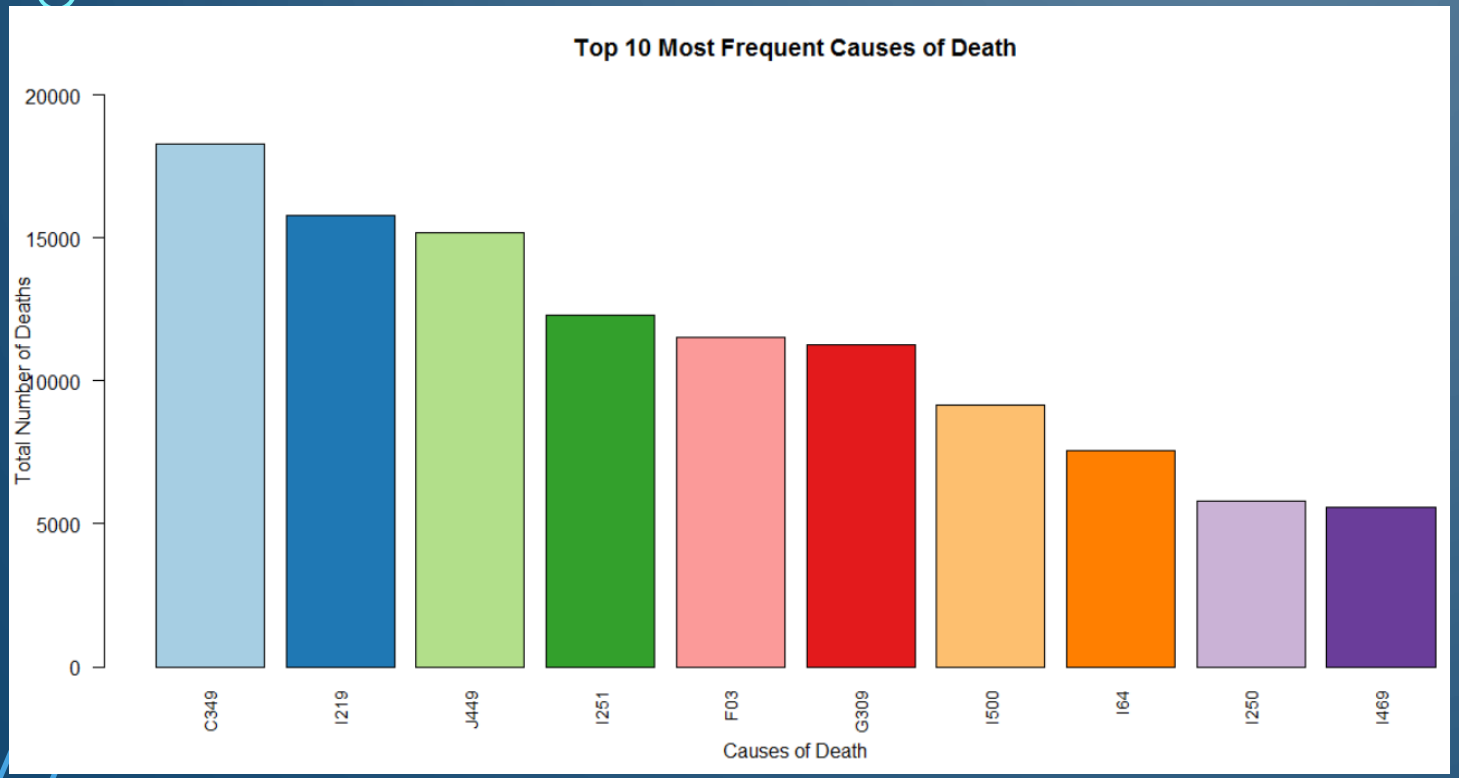
STATISTICAL ANALYSIS

- Pattern of Manner of Death across several Races



STATISTICAL ANALYSIS

- Frequent cause of Natural Death



CauseOfDeath	count	Description
C349	18278	Malignant neoplasm of bronchus and lung
I219	15770	Acute myocardial infarction
J449	15170	Other chronic obstructive pulmonary disease
I251	12316	Chronic ischaemic heart disease
F03	11513	Unspecified dementia
G309	11275	Alzheimer disease
I500	9149	Heart failure
I64	7549	Stroke, not specified as haemorrhage or infarction
I250	5804	Chronic ischaemic heart disease
I469	5565	Cardiac arrest

STATISTICAL ANALYSIS - SUICIDE

- Who likely take their own lives – Male or Female?

```
> suicide.ratio
  Sex  Cases     Mean      Std       pct
1   F 145004 73.29274 19.07906 0.4833467
2   M 154996 68.13377 18.35433 0.5166533
```

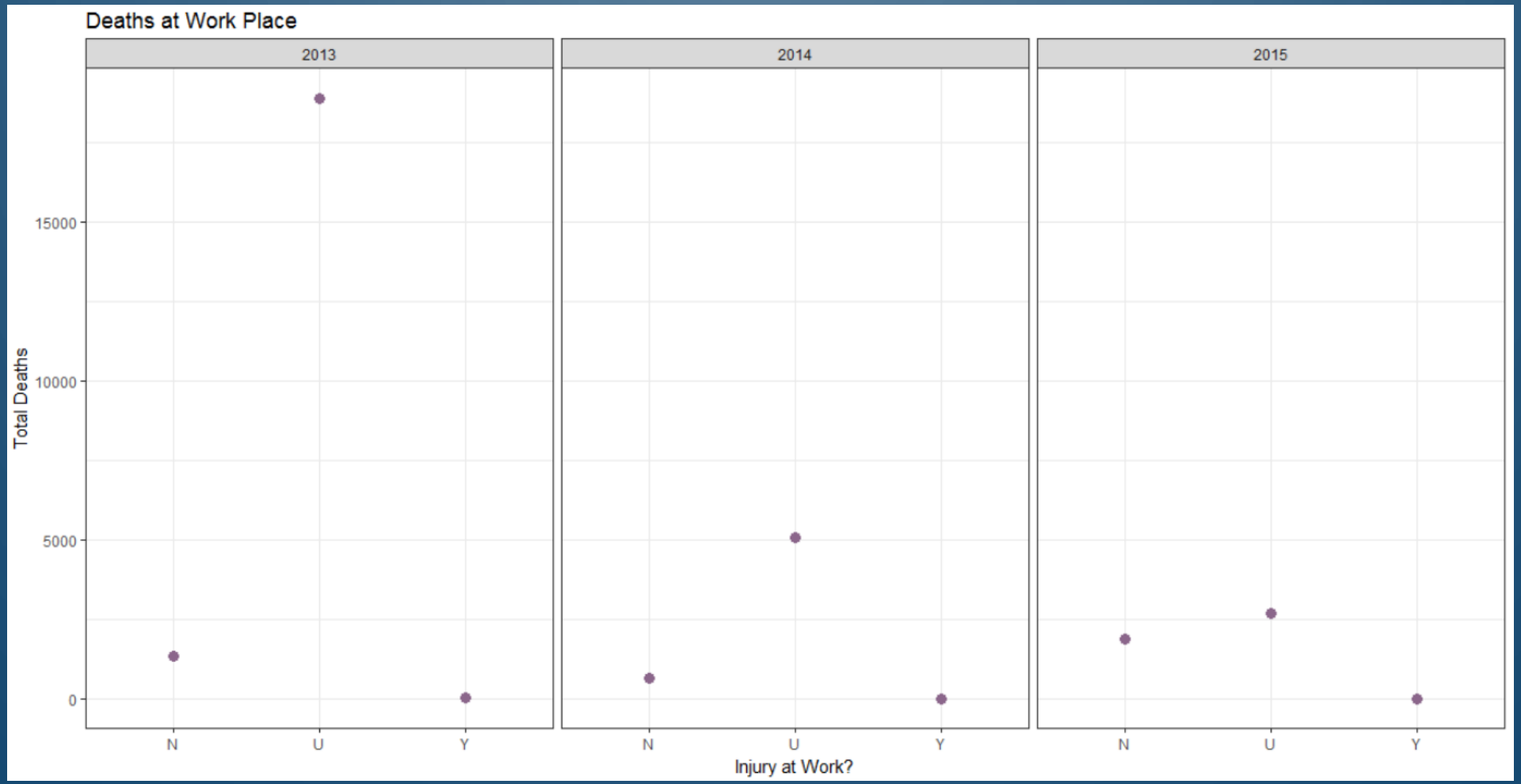
- Average Life Expectancy
  - Male – 68
  - Female – 73
- Percentage of Suicide deaths
  - Male – 52%
  - Female – 48%

STATISTICAL ANALYSIS - SUICIDE

- Is Marital Status has effect on suicidal cases?

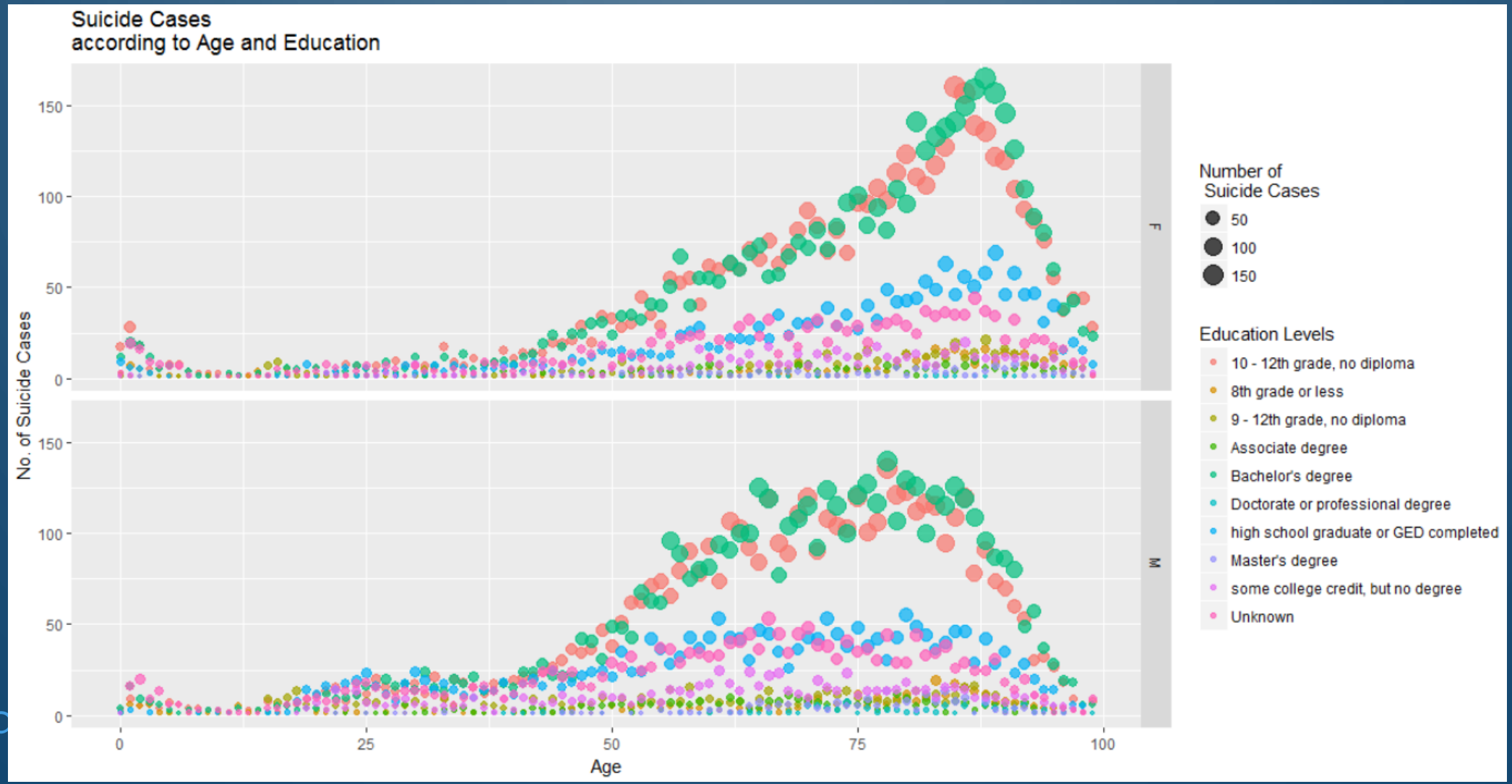
STATISTICAL ANALYSIS - SUICIDE

- Is Work pressure a reason for majority of suicides?



STATISTICAL ANALYSIS - SUICIDE

- Suicide analysis based on education, Sex and Age



MODEL FITTING - SUICIDE

- Preprocessing – Addition of Suicide column
  - 1 – Yes
  - 0 – No

- Model 1 : Linear Regression
- Model 2 : Logistic Linear model
- Model 3 : Linear Regression with more variables
- Model 4 : Linear Regression with Interaction variables

MODEL FITTING - SUICIDE

- Model Comparison

Model	Predictors	R square	Adj R square	AIC	BIC
		(increase)	increase)	(decrease)	(decrease)
Model 1	Education+Sex+Age+MaritalStatus+Race	0.34	0.33	7656.312	7858.941
Model 2	Education + Sex + Age + MaritalStatus + Race + MonthOfDeath	0.55	0.49	7375.6	7627.013
Model 3	ResidentStatus+Education+Sex+Age+CauseOfDeath+ MaritalStatus+ Race+MonthOfDeath+DayOfWeekOfDeath	0.53	0.49	6305.636	12007.19
Model 4	ResidentStatus+Education+Sex+Age+ CauseOfDeath+MaritalStatus+ Race+MonthOfDeath+ DayOfWeekOfDeath+Education*Age	0.54	0.49	6267.019	12031.46

MODEL FITTING - SUICIDE

- Anova analysis (Analysis of Variance)

```
> anova(model2, test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: Suicide

Terms added sequentially (first to last)


             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                          7999    10436.1
Education     9  2842.25      7990     7593.9 < 2.2e-16 ***
Sex           1     9.54      7989     7584.4  0.002013 **
Age           1    89.25      7988     7495.1 < 2.2e-16 ***
MaritalStatus 4     9.25      7984     7485.9  0.055188 .
Race         12    37.27      7972     7448.6  0.000202 ***
MonthOfDeath  8   145.11      7964     7303.5 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

MODEL FITTING - SUICIDE

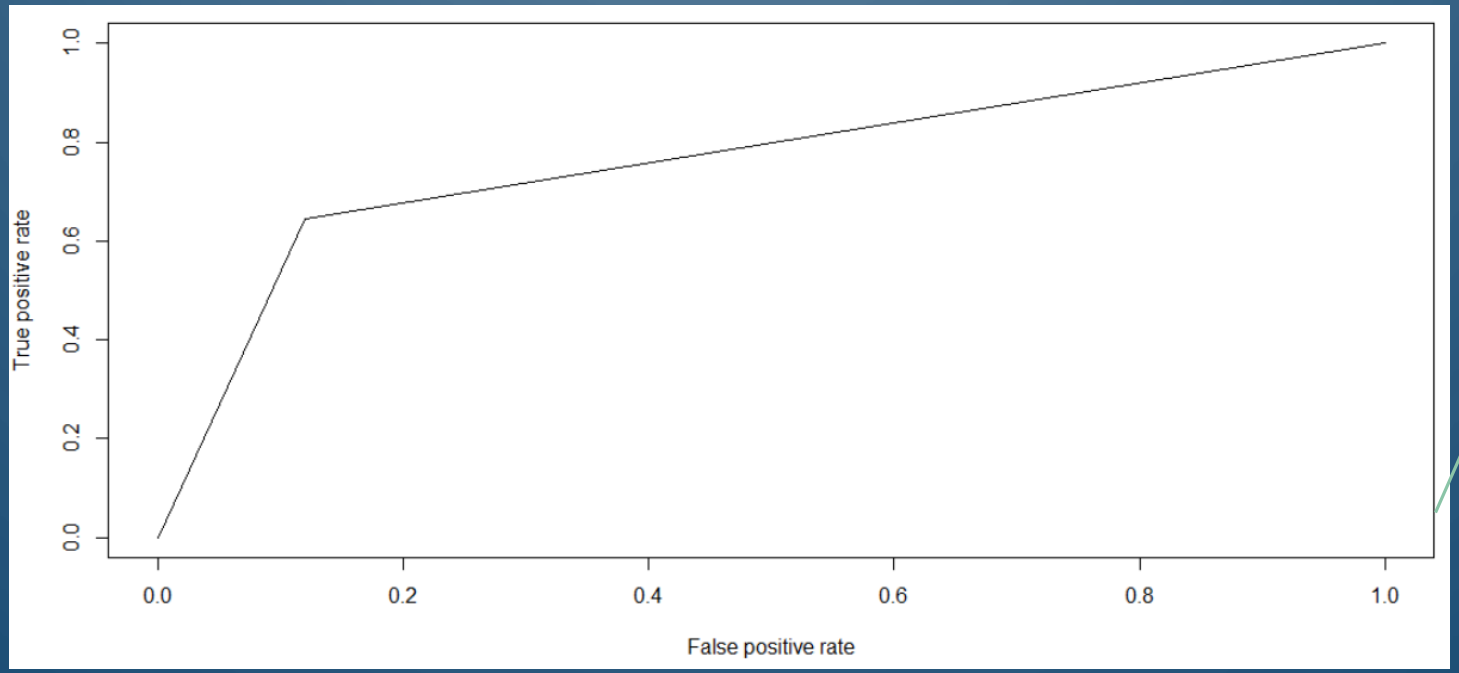
- Passing Test data through model2
- Accuracy of the model 2 – 80%

```
> fitted.results <- predict(model2,newdata=test,type='response')
> fitted.results <- ifelse(fitted.results > 0.5,1,0)
> misClasificError <- mean(fitted.results != test$Suicide)
> print(paste('Accuracy',1-misClasificError))
[1] "Accuracy 0.802024746906637"
```

MODEL FITTING - SUICIDE

- ROC Curve to calculate area under the curve – best for binary classifier

```
> pr <- prediction(fitted.results, test$Suicide)
> prf <- performance(pr, measure = "tpr", x.measure = "fpr")
> plot(prf)
```

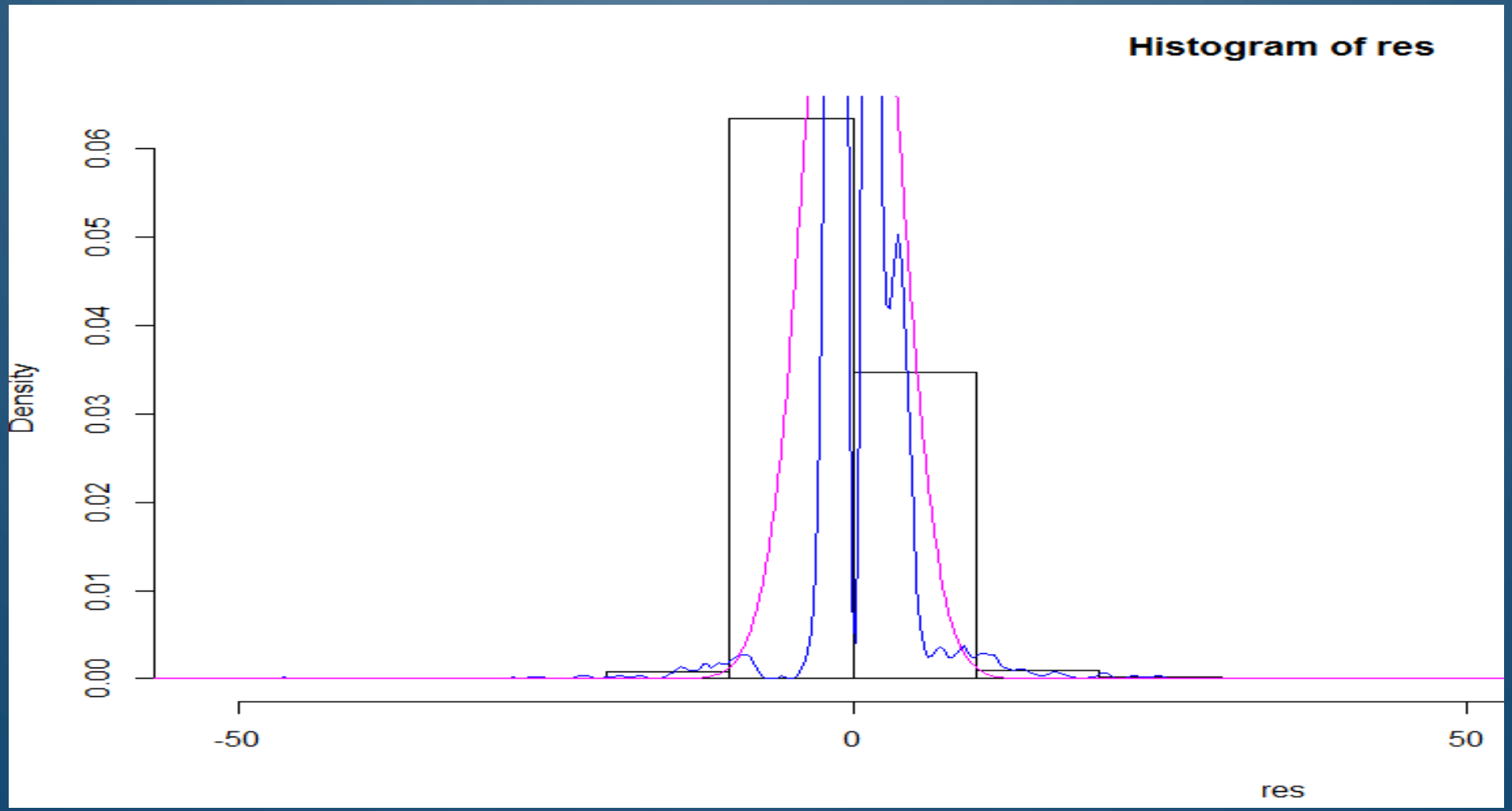


Curve closer to 1
->
good predictive ability of model

```
> auc <- performance(pr, measure = "auc")
> auc <- auc@y.values[[1]]
> auc
[1] 0.76277
```

MODEL FITTING - SUICIDE

- Residual density plot (Blue) – against normal distribution (Pink)



WHAT WE'VE ANALYZED…

1. People from age group 70 to 80, especially Widowed, are more likely to die - Obvious!

2. Winter season is reported with high death rate

3. High school graduates are reported with more number of deaths

4. Primary manner of death observed is - "Natural" with high number of deaths caused by Cancer

5. Suicide trends are observed increasing with year

6. Life expectancy for Male is less, as they are observed committing suicides more frequently than females