R Notebook

Code

#1

Hide

```
library(dplyr)
library(tidyr)
library(ggplot2)
library(RColorBrewer)
```

Hide

```
Stroke_Date<- read.csv(file.choose())
```

Hide

```
Stroke_Date
```

id <int>	gen… <chr>	age <dbl>	hypertension <int>	heart_disease <int>	ever_married <chr>	work_type <chr>	Residence_type <chr>
9046	Male	67.00	0	1	Yes	Private	Urban
51676	Female	61.00	0	0	Yes	Self-employed	Rural
31112	Male	80.00	0	1	Yes	Private	Rural
60182	Female	49.00	0	0	Yes	Private	Urban
1665	Female	79.00	1	0	Yes	Self-employed	Rural
56669	Male	81.00	0	0	Yes	Private	Urban
53882	Male	74.00	1	1	Yes	Private	Rural
10434	Female	69.00	0	0	No	Private	Urban
27419	Female	59.00	0	0	Yes	Private	Rural
60491	Female	78.00	0	0	Yes	Private	Urban

1-10 of 5,110 rows | 1-8 of 12 columns Previous **1** 2 3 4 5 6 … 100 Next

ID is not a meaningful quantitative variable because its values do not have any patterns in them. Here it is being used as a unique identifier.

I chose to make hypertension, heart_disease and stroke columns into logical variables as they indicate whether or not someone has these conditions. As the BMI column contains values with decimals in them, I stored it as a double. As the Smoking_status column is a cateogrical variable, I stored it as a factor.

Hide

```
Stroke_Date$hypertension<-as.logical(Stroke_Date$hypertension)
Stroke_Date$heart_disease<-as.logical(Stroke_Date$heart_disease)
Stroke_Date$stroke<-as.logical(Stroke_Date$stroke)
Stroke_Date$smoking_status<-as.factor(Stroke_Date$smoking_status)
Stroke_Date$bmi<-as.double(Stroke_Date$bmi)
```

```
NAs introduced by coercion
```

Hide

```
Stroke_Date
```

id <int>	gen… <chr>	age <dbl>	hypertension <lgl>	heart_disease <lgl>	ever_married <chr>	work_type <chr>	Residence_type <chr>
9046	Male	67.00	FALSE	TRUE	Yes	Private	Urban
51676	Female	61.00	FALSE	FALSE	Yes	Self-employed	Rural
31112	Male	80.00	FALSE	TRUE	Yes	Private	Rural
60182	Female	49.00	FALSE	FALSE	Yes	Private	Urban
1665	Female	79.00	TRUE	FALSE	Yes	Self-employed	Rural
56669	Male	81.00	FALSE	FALSE	Yes	Private	Urban
53882	Male	74.00	TRUE	TRUE	Yes	Private	Rural
10434	Female	69.00	FALSE	FALSE	No	Private	Urban
27419	Female	59.00	FALSE	FALSE	Yes	Private	Rural
60491	Female	78.00	FALSE	FALSE	Yes	Private	Urban

1-10 of 5,110 rows | 1-8 of 12 columns Previous **1** 2 3 4 5 6 … 100 Next

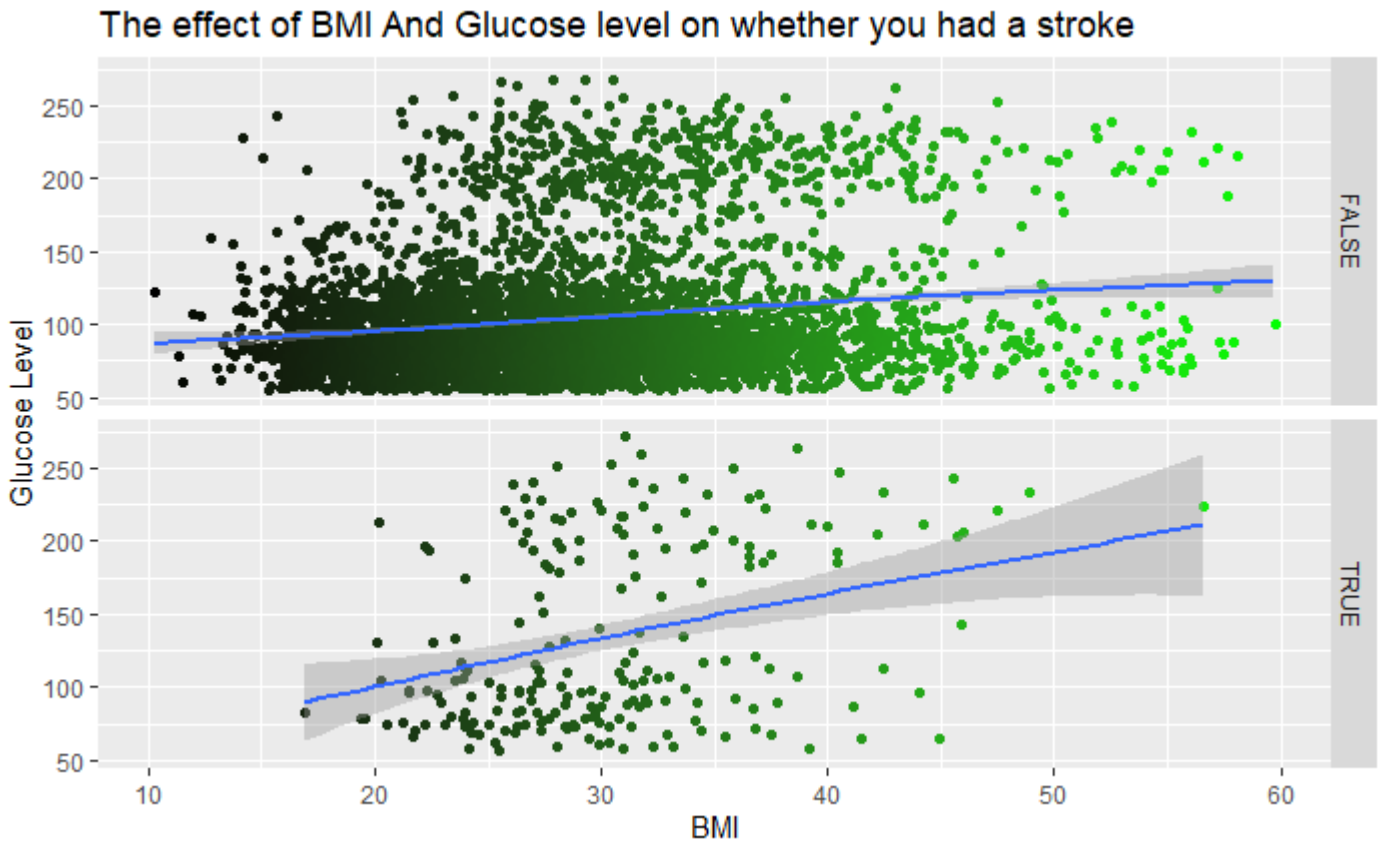
#2

Hide

```
Stroke_Date<-Stroke_Date%>%filter(bmi<60)
```

Hide

```
ggplot(Stroke_Date,aes(x=bmi,y=avg_glucose_level,color=bmi))+geom_jitter()+facet_grid(stroke~.)+
geom_smooth()+scale_color_gradient(low="black",high="green")+labs(title="The effect of BMI And G
lucose level on whether you had a stroke",x="BMI" ,y="Glucose Level")+ theme(legend.position =
"none")
```



This graph shows that there is a positive correlation between the BMI and the average Glucose level in individual who have and haven't had a stroke. This figure is showing that individuals who have had a stroke are more likely to have higher glucose level.

#3

Hide

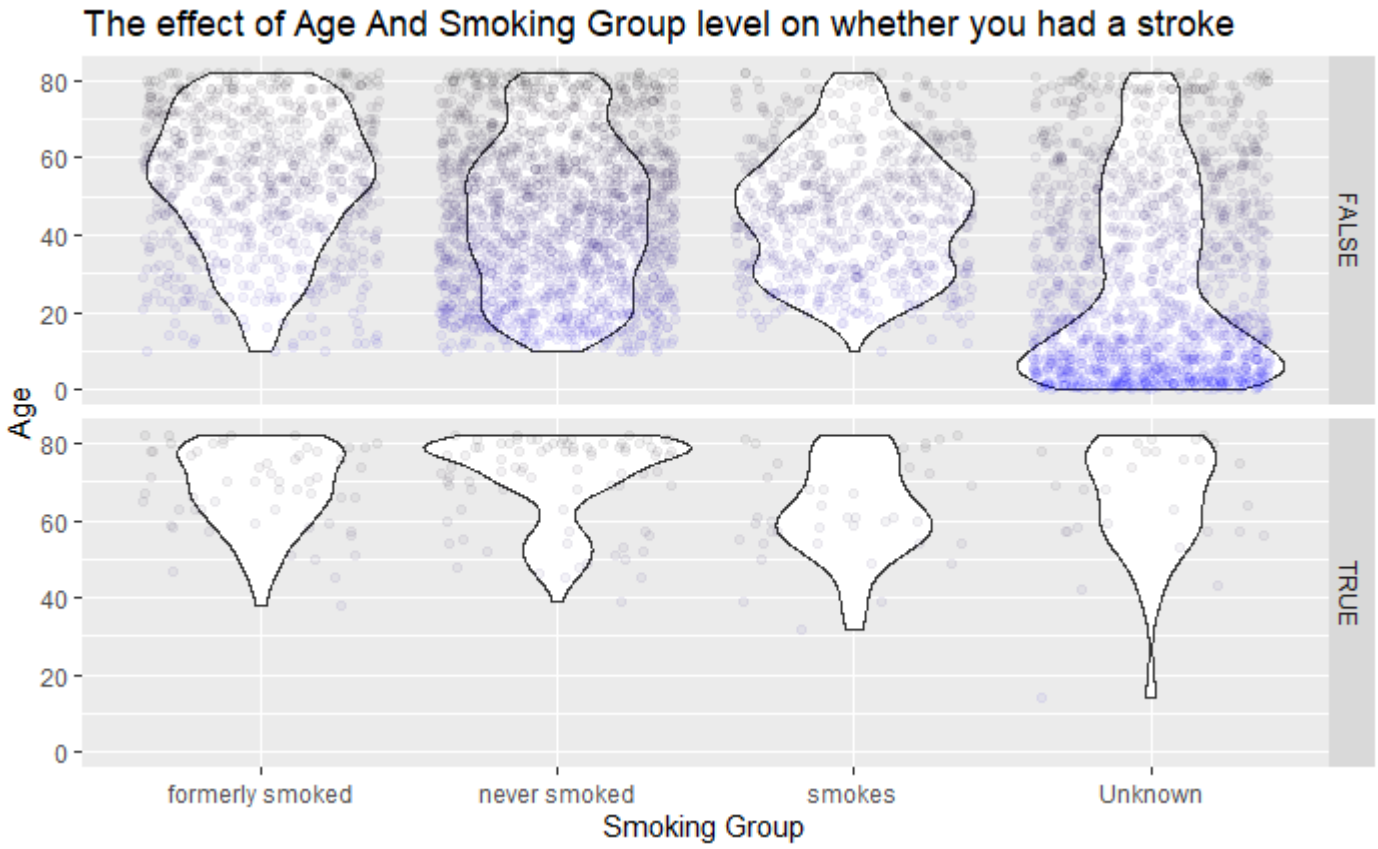
```
Stroke_Date
```

id <int>	gen… <chr>	age <dbl>	hypertension <lgl>	heart_disease <lgl>	ever_married <chr>	work_type <chr>	Residence_type <chr>
9046	Male	67.00	FALSE	TRUE	Yes	Private	Urban
31112	Male	80.00	FALSE	TRUE	Yes	Private	Rural
60182	Female	49.00	FALSE	FALSE	Yes	Private	Urban
1665	Female	79.00	TRUE	FALSE	Yes	Self-employed	Rural
56669	Male	81.00	FALSE	FALSE	Yes	Private	Urban
53882	Male	74.00	TRUE	TRUE	Yes	Private	Rural
10434	Female	69.00	FALSE	FALSE	No	Private	Urban
60491	Female	78.00	FALSE	FALSE	Yes	Private	Urban
12109	Female	81.00	TRUE	FALSE	Yes	Private	Rural
12095	Female	61.00	FALSE	TRUE	Yes	Govt_job	Rural

1-10 of 4,896 rows | 1-8 of 12 columns Previous **1** 2 3 4 5 6 … 100 Next

Hide

```
ggplot(Stroke_Date,aes(x=smoking_status,y=age,color=age))+geom_violin()+facet_grid(stroke~.)+geo
m_jitter(alpha=.05)+scale_color_gradient(low="Blue",high="black")+ theme(legend.position = "non
e")+labs(title="The effect of Age And Smoking Group level on whether you had a stroke",x="Smokin
g Group" ,y="Age")
```



As we can see from the above graph, people who are former smokers have the largest proportion of people 60 or older. This makes sense as it takes time to give up smoking. Conversely, for people whose smoking status is unknown, this group has the largest proportion of people who are younger than 10 years old. This makes sense as the people below 10 may not have been asked if they smoke as they are very young. The graph shows that there is a lot more data for people who have never had strokes than those who have had. This is likely cause the majority of participant in the data set did not have a stroke. The people who have had a stroke tend to be older than those who have not for all smoking groups.