stroke

Can we predict the stroke? Looking into factors that may contribute to getting a stroke

Research question

Stroke is one of the **leading causes of death and disability** in the UK and in the world. According to UK statistics **100,000 people** have stroke each year. Although anyone can have a stroke at any age, certain medical conditions and lifestyle factors may increase the risk of having a stroke. However, it is not known which one of these may have stronger influence than the other. The main purpose of this visualisation is to investigate if any of the factors occurs more often than the other in stroke patients and if there is any correlation between body mass index and blood glucose levels and stroke occurrence.

Data origins

The data used in this visualisation project has been downloaded from Kaggle which is a data repository of community published data & code.

The data set was meant to be used to predict the likelihood of stroke occurrence. It consists over 5000 patient records that include:

- gender,
- age,
- medical conditions:
  - *hypertension*,
  - *heart disease*

Additionally collected data include:

- body mass index,
- average glucose level,
- work type:
  - *private*,
  - *self-employment*,
  - *government job*,
  - *home work due to having children*,
- residence type:
  - *urban*
  - *rural*
- smoking status:
  - *never*,
  - *formerly smoking*,

  - *smokes*,
- whether patient had a stroke or not.

Unfortunately the source of this data is classified as confidential.

Please find the link to the data on Kaggle.com.

Initial data checks and preparation

```
#importing of the data
df <- read.csv(here("data", "stroke.csv"))
head(df, n=3)  #displaying the top of the data frame, only 3 rows
```

```
##      id gender age hypertension heart_disease ever_married     work_type
## 1  9046   Male  67            0             1          Yes       Private
## 2 51676 Female  61            0             0          Yes Self-employed
## 3 31112   Male  80            0             1          Yes       Private
##   Residence_type avg_glucose_level  bmi  smoking_status stroke
## 1          Urban            228.69 36.6 formerly smoked      1
## 2          Rural            202.21   NA    never smoked      1
## 3          Rural            105.92 32.5    never smoked      1
```

```
#unification - changing binary values (1,0) to Yes,No
#as the stroke, hypertension,heart disease columns had 1,0 but ever_married had Yes,No
df$stroke <- df$stroke %>% recode('1' = "Yes", '0' = "No")
df$hypertension <- df$hypertension %>% recode('1' = "Yes", '0' = "No")
df$heart_disease <- df$heart_disease %>% recode('1' = "Yes", '0' = "No")
```

Foreword The original data set contains 5110 rows. Unfortunately 201 had to be removed as had blanks for BMI. Nevertheless, although BMI was missing, there were other information in the rows, that I did not want to lose. Therefore I have first filtered out and created new dataset (df_stroke) for the stroke group only that was used only for the Visualisation 2. It does contain blank BMI rows, however it's not important as BMI is not included on that graph.

However, for the purpose of Visualisation 1, I have removed the blank BMI rows which left me with 4909 rows (209 stroke and 4700 no stroke). I have sampled the no stroke group for 209 random rows to match with the stroke group. Visualisation 1 is based on the group of 418 individuals (209 stroke, 209 no stroke).

```
#setting dataset for Visualisation 2 with only stroke patients
df_stroke <- filter(df,stroke =="Yes")

#checking how many blanks and in which table columns
colSums(is.na(df))
```

```
##                id            gender               age      hypertension
##                 0                 0                 0                 0
##     heart_disease      ever_married         work_type    Residence_type
##                 0                 0                 0                 0
## avg_glucose_level               bmi    smoking_status            stroke
##                 0               201                 0                 0
```

```r
#deleting empty BMI rows and setting new clean data set
df_new <- df[!(is.na(df$bmi) | df$bmi==""), ]

#separate data sets for stroke and no stroke
df_new_no_stroke <- filter(df_new, stroke == "No")
df_new_stroke <- filter(df_new, stroke == "Yes")

#randomly choosing 209 no stroke rows to have equal number as with the stroke group
df_new_no_stroke_sample <- df_new_no_stroke %>% sample_n(209, replace = FALSE, prob = NULL)

#combining new datasets for Vis 1 into new data set of 418 (209 each group) values
df_new_sample <- rbind(df_new_stroke, df_new_no_stroke_sample)

#while inspecting the data notice gender as other so checking how many and removing
nrow(df_new_sample[df_new_sample$gender == "Other",])
```

```
## [1] 0
```

```r
df_new_sample <- filter(df_new_sample, gender !="Other")

#total gender count in the sample
males_total <- nrow(df_new_sample[df_new_sample$gender == "Female",])
females_total <- nrow(df_new_sample[df_new_sample$gender == "Male",])
```

Average glucose levels and BMI over lifespan and stroke occurence

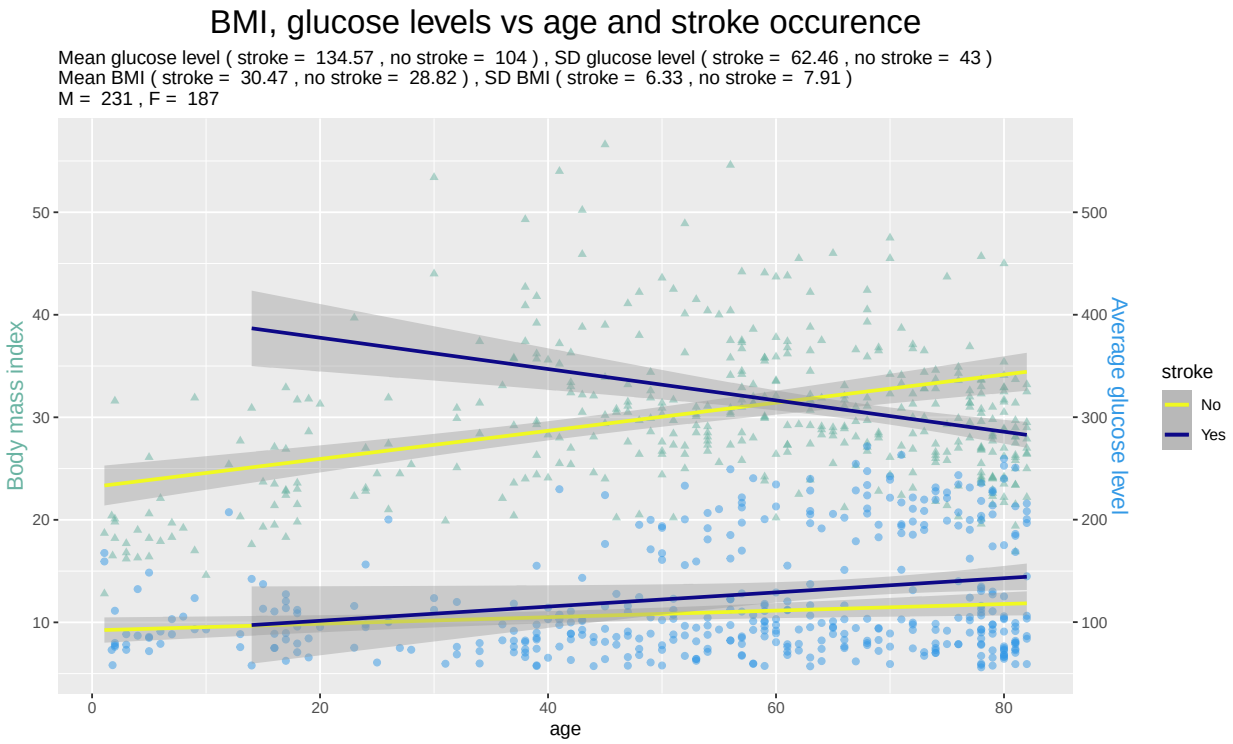
```r
#means and sd and saving as data frame
means_stroke <- df_new_stroke %>% summarise_if(is.numeric, mean)
means_no_stroke <- df_new_no_stroke %>% summarise_if(is.numeric, mean)
sd_stroke <- df_new_stroke %>% summarise_if(is.numeric, sd)
sd_no_stroke <- df_new_no_stroke %>% summarise_if(is.numeric, sd)

#rounding to 2 decimal places
means_stroke <- round(means_stroke, digits = 2)
means_no_stroke <-round(means_no_stroke, digits =2)
sd_stroke <- round(sd_stroke, digits = 2)
sd_no_stroke <- round(sd_no_stroke, digits = 2)
```

Preparation of the data for the first visualisation

```r
coeff <- 10

#colour spec to use in the visualisation
bmi_color <- "#69b3a2"
avg_glucose_color <- rgb(0.2, 0.6, 0.9, 1)
```

```r
#plot vis1
ggplot(df_new_sample, aes(x=age, col=stroke)) +

  geom_point( aes(y=bmi), color=bmi_color, shape=17, alpha=0.5) +
  geom_point( aes(y=avg_glucose_level /coeff), color=avg_glucose_color, shape = 19, alpha=0.5) +
  scale_colour_viridis_d(option = "C", direction = -1) + #colorblind palette

  geom_smooth(aes(y=bmi), method = "lm")+
  geom_smooth(aes(y=avg_glucose_level /coeff), method = "lm") +
  scale_y_continuous(
    name = "Body mass index",
    sec.axis = sec_axis(~.*coeff, name="Average glucose level")
  ) +
  theme(
    axis.title.y = element_text(color = bmi_color, size=13),
    axis.title.y.right = element_text(color = avg_glucose_color, size=13),
        plot.title = element_text(hjust = 0.5, size = 19)
    ) +
  labs(title="BMI, glucose levels vs age and stroke occurence",
       subtitle = paste(
       "Mean glucose level ( stroke = ",means_stroke$avg_glucose_level,
       ", no stroke = ", means_no_stroke$avg_glucose_level,")",
       ", SD glucose level ( stroke = ",sd_stroke$avg_glucose_level,
       ", no stroke = ", sd_no_stroke$avg_glucose_level,
       ")\nMean BMI ( stroke = ",means_stroke$bmi, ", no stroke = ", means_no_stroke$bmi,")",
       ", SD BMI ( stroke = ",sd_stroke$bmi, ", no stroke = ", sd_no_stroke$bmi,")",
       "\nM = ", males_total, ", F = ", females_total))
```

Visualisation 1: Body mass index average glucose level vs age and stroke occurence




```
ggsave(here("plots", "visualisation_1_bmi_glucose_stroke.png"))
```

Commentary The graph above allows to conclude, that glucose levels not only are increasing with age but also that individuals that suffered from stroke have higher glucose levels (mean avg glucose levels stroke = 134.57 vs no stroke = 104). As diabetes is a known risk for cardiovascular disease, which can lead to a stroke, the observed tendency is plausible. Nevertheless, body mass index opposite to the glucose levels, is not significantly higher in people that had stroke (mean bmi stroke = 30.47 vs no stroke 28.82). Although, it can be observed that BMI increases with age, which could be a result of decrease in physical activity with age.

Stroke factors overview based on people who suffered from stroke

```
#gender count of stroke patients
males_stroke_total <- nrow(df_stroke[df_stroke$gender == "Male",])
females_stroke_total <- nrow(df_stroke[df_stroke$gender == "Female",])

#checking how many patients had stroke and how many did not in the stroke group
n_stroke <- nrow(df[df$stroke == "Yes",])
n_no_stroke <-nrow(df[df$stroke == "No",])

#rearranging columns so can change data from wide to long
df_stroke <- df_stroke %>% relocate(smoking_status, .before = avg_glucose_level)
df_stroke <- df_stroke %>% relocate(gender, .before = smoking_status)

#changing from wide to long type data
```

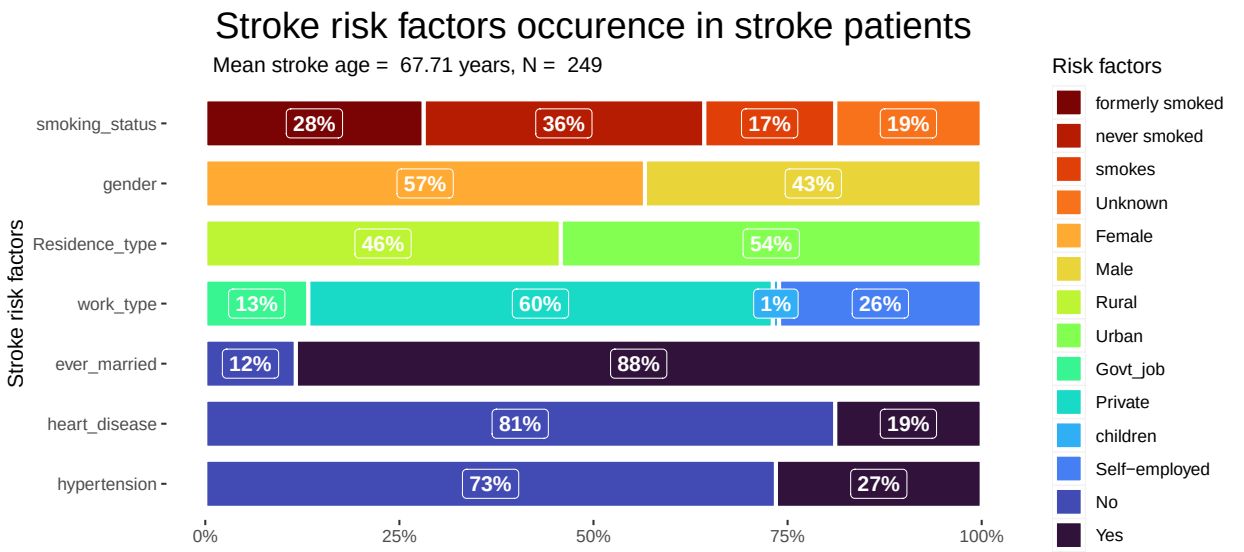
```
#so chosen columns are now as rows and can use them in bar chart below
df_stroke$id <- factor(df_stroke$id)
df_stroke <- gather(df_stroke, condition, measurement,
                    hypertension:smoking_status, factor_key = TRUE)
```

Preparing the data for the second visualisation

```
#stacked barchart

df_stroke$measurement <- factor(df_stroke$measurement, levels = c("Unknown","smokes","never smoked", "f
               "Male", "Female", "Urban","Rural", "Self-employed", "children",
               "Private","Govt_job","Yes", "No"))

p2 <- ggplot(df_stroke,aes(condition, fill=measurement))
  p2 + geom_bar(position = 'fill', stat = "count",
    width = 0.8,
    colour = "white",
    size = 1) +
  scale_fill_viridis_d(
    option="H",
    direction = -1,
    # arranging legend order to correspond to variables order
    limits = c("formerly smoked", "never smoked", "smokes", "Unknown",
               "Female", "Male", "Rural","Urban", "Govt_job",
               "Private", "children", "Self-employed", "No", "Yes")) +
  scale_colour_manual(values=c("#FFFFFF"))+
  geom_label(data = . %>%
               group_by(condition, measurement) %>%
               tally() %>%
               mutate(p = n / sum(n)) %>%
               ungroup(),
             aes(y = p, label = scales::percent(p, accuracy = 1),
    fontface = "bold", color = "white"),
             position = position_stack(vjust = 0.5),
             show.legend = FALSE) +
   scale_y_continuous(labels= scales::percent) +
  theme(axis.title.x=element_blank(),
        panel.background = element_blank(),
        plot.title = element_text(hjust = 0.5, size = 19),
        plot.subtitle = element_text(hjust = 0.1))+
  #title and subtitle of the chart
  labs(
    title="Stroke risk factors occurence in stroke patients",
    subtitle = paste("Mean stroke age = ",means_stroke$age,"years,", "N = ", n_stroke),
    x = "Stroke risk factors") +
  coord_flip() +
 guides(
  fill = guide_legend(
    title = "Risk factors",
    override.aes = aes(label = "")
```

```
  )
)
```

Visualisation 2: Stroke risk factors occurence in stroke patients




```
ggsave(here("plots", "visualisation_2_risk_factors.png"))
```

Commentary The purpose of this chart was to look at only the stroke patients (N= 249) risk factors to see if any factor stands out more due to occurring in more individuals. Surprisingly majority of the individuals that got stroke were married. Although correlation does not cause causation, we could point a stress as contributing factor. Nevertheless, we can see slight differences for gender (F = 141 vs M = 108) and work type (Private = 149) and leaving in urban area (N = 135). Smoking status data could be interesting to evaluate as its known for causing cardiovascular diseases however 47 records did not contain this information. Beside the marriage factor, unfortunately none of these factors occurs often enough in stroke patients to draw any major conclusions. Surprisingly and opposite to scientific research results, the heart disease and hypertension were not present in many patients that suffered from stroke.

Summary and Discussion

Initially when started this project, I was a bit confused as what to do with the data. I wanted to use as much of data in the set as possible to create my visualisation. Along the way, I have tried different plots, tried using animation, but decided that it does not add any benefit to what I want to present. Nevertheless, once progressed, understood more and more how R works, so it became easier and every day had more ideas and was eager to try if they work. My self-learning got to the point that could have quickly identify what is wrong and how to make it work. Substantial amount of time I have spent on re-reading the code and trying to find simpler and less space-consuming code that works in the same way. I have truly enjoyed this project and will definitely use R going forward for my work instead of Excel.

Caveats:

1. Missing values for smoking status (47) so could not assess properly the impact of smoking

2. Small amount of stroke patients (249), it would be great to have equal size sample as no stroke (4861)
3. There is nothing about the sample ethnicity, location, year, that would be a great addition for more thorough investigation on the subject.
4. It was a small data set with limited possibilities.

Published: Link to github repo