HW 03 - Road traffic accidents

In this assignment we'll look at traffic accidents in Edinburgh. The data are made available online by the UK Government. It covers all recorded accidents in Edinburgh in 2018 and some of the variables were modified for the purposes of this assignment.

Packages

We'll use the **tidyverse** package for much of the data wrangling and visualisation and the data lives in the **dsbox** package. These packages are already installed for you. You can load them by running the following in your Console:

```
library(tidyverse)
library(dsbox)
library(ggridges)
```

Data

The data can be found in the **dsbox** package, and it's called `accidents`. Since the dataset is distributed with the package, we don't need to load it separately; it becomes available to us when we load the package. You can find out more about the dataset by inspecting its documentation, which you can access by running `?accidents` in the Console or using the Help menu in RStudio to search for `accidents`. You can also find this information here.

Exercises

1. How many observations (rows) does the dataset have? Instead of hard coding the number in your answer, use inline code.

The dataset has 768 rows.

2. Run `View(accidents)` in your Console to view the data in the data viewer. What

does each row in the dataset represent?

Each row represent and accident, first beggining with the accident ID and then it has a lot of information about the accident such as where it happened, data, how many vehicles, time, weather etc.

3. Answer the following questions and include at least one data visualization. Hint: a density plot may be useful for this one

1. When are accidents most likely to occur?
   Accidents are most likely to happen between 10 AM and 8 PM

2. Does it vary by severity?

Yes, it does, between 10 AM and 8 PM we see that there are more accidents with severity labeled slight, there are less with severity labeled moderate and those labeled fatal are very few

3. Are there any differences between weekends and weekdays?
   No, as we can see in the heat map between slight and moderate accidents there isnt much a difference of between weekdays and weekends, where as for fatal accidents we dont see any accident over the weekend

```
accidents_hour_grouped <- accidents %>%
  group_by(day_of_week, time = as.numeric(substr(time, 1, 2))) %>%
  summarise(count = n())
```

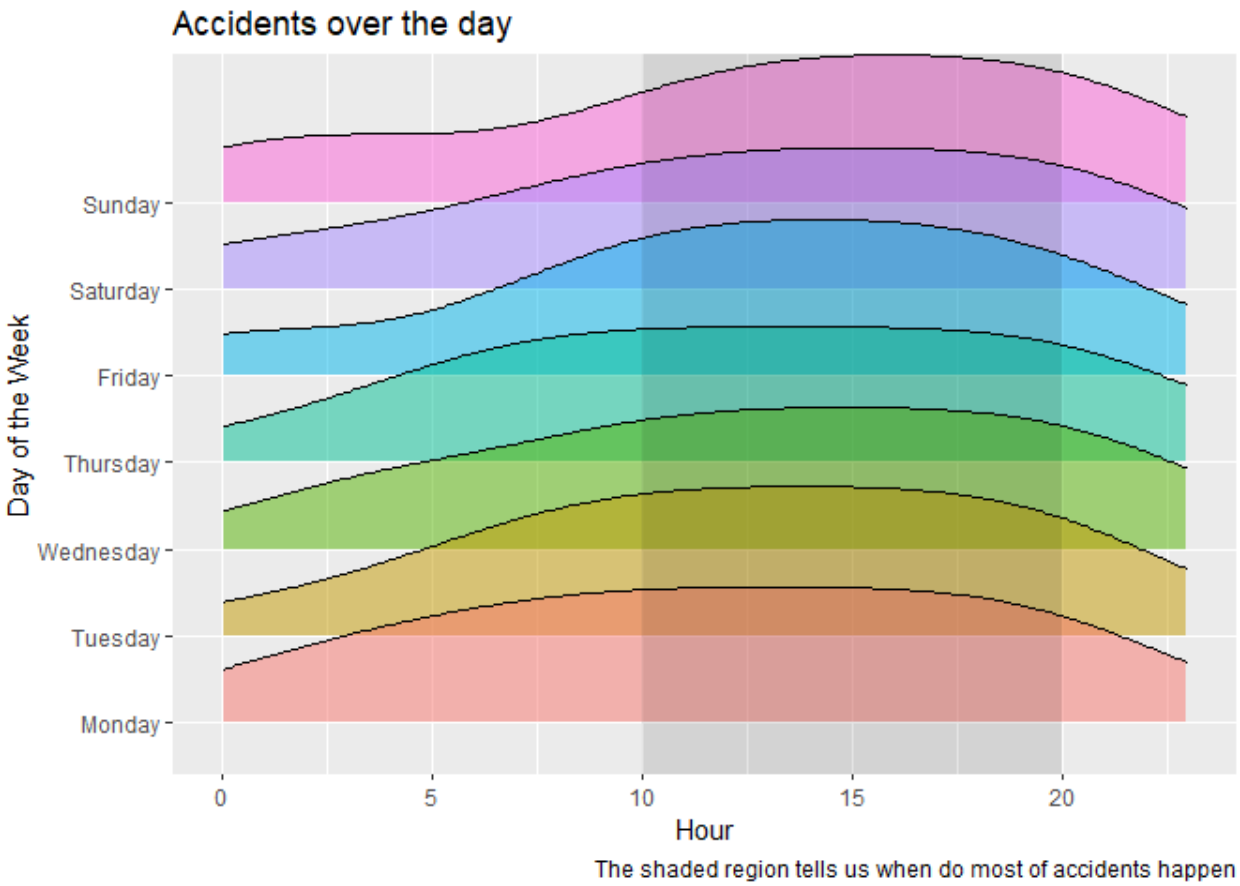
```
## `summarise()` has regrouped the output.
## ℹ Summaries were computed grouped by day_of_week and time.
## ℹ Output is grouped by day_of_week.
## ℹ Use `summarise(.groups = "drop_last")` to silence this message.
## ℹ Use `summarise(.by = c(day_of_week, time))` for per-operation grouping
##   (`?dplyr::dplyr_by`) instead.
```

```
#this chunk is self explanatory, fist i group the data by the day of the week and

ggplot(accidents_hour_grouped, aes(x = time, y = day_of_week, fill = day_of_week))
  geom_density_ridges(alpha = 0.5) +
  scale_x_continuous(limits = c(0,23))+
  labs(
    title = "Accidents over the day",
```

```
    x = "Hour",
    y = "Day of the Week",
    caption = "The shaded region tells us when do most of accidents happen"
  ) +
  annotate("rect", xmin = 10, xmax = 20, ymin = -Inf, ymax = Inf, alpha = 0.1, fil
  theme(legend.position = "none")
```

```
## Picking joint bandwidth of 3.21
```

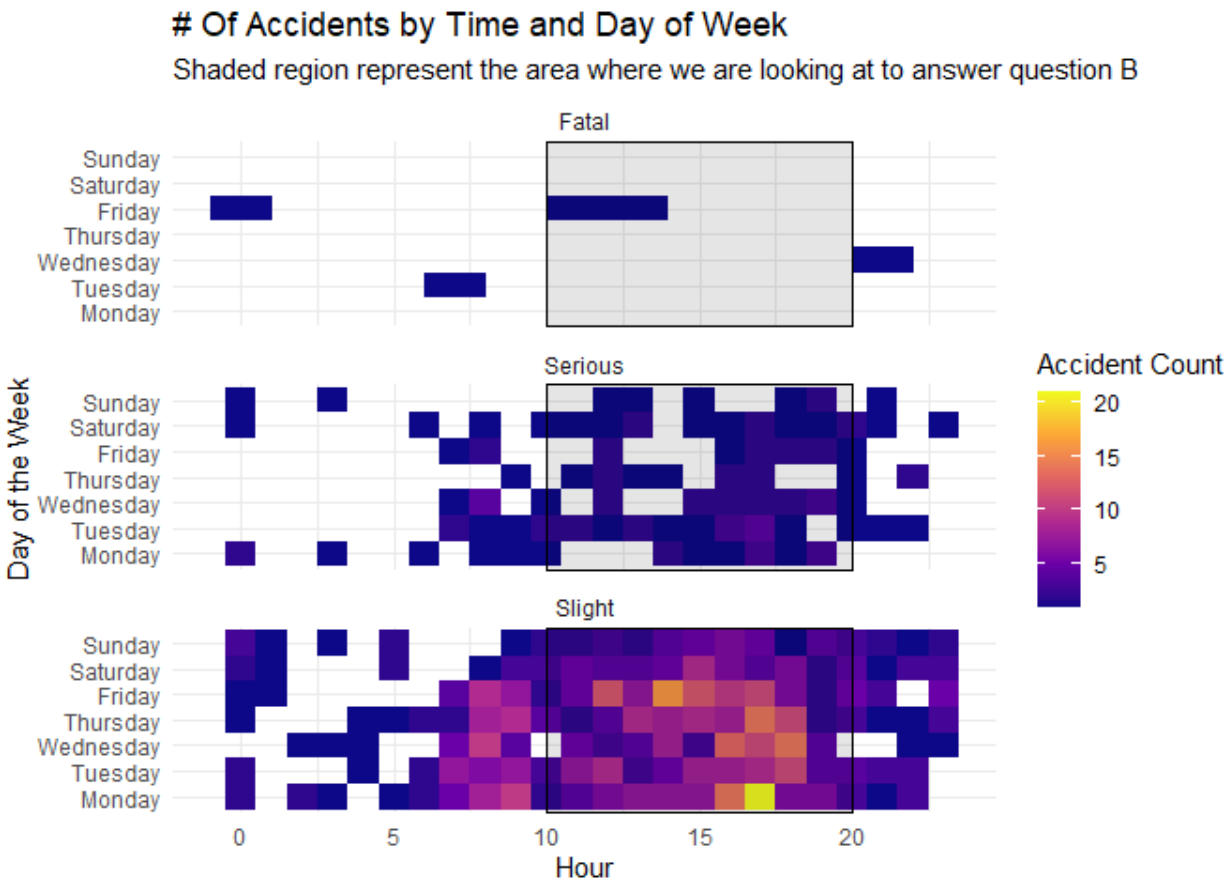



```
# i chose a density ridge to plot the data because its easy to read the trend of a
```

```
accidents_severity <- accidents %>%
  count(day_of_week,time = as.numeric(substr(time, 1, 2)), severity)

ggplot(accidents_severity, aes(x = time, y = day_of_week, fill = n)) +
  geom_tile() +
  facet_wrap(~ severity, ncol  = 1) + # faceting by severity
  scale_fill_viridis_c(option = "C", name = "Accident Count") + #giving the heat m
  annotate("rect", xmin = 10, xmax = 20, ymin = -Inf, ymax = Inf, alpha = 0.1, fil
  labs(
```

```
    title = "# Of Accidents by Time and Day of Week",
    x = "Hour",
    y = "Day of the Week",
    subtitle = "Shaded region represent the area where we are looking at to answer
  ) +
  theme_minimal()
```



4. Propose and answer at least one more question from this data set. And create another data visualisation based on these data and interpret it for your answer. You can choose any variables and any type of visualisation you like, but it must have at least three variables, e.g. a scatterplot of x vs. y isn't enough, but if points are coloured by z, that's fine.

1. What is the question you are trying to answer?
   Does the severity of the crash increase as the speed limit increases?
2. What is the answer?
   No. The scatterplot plots the data by given longitude and latitudal data, and then the higher the speed limit, the more visible the point is, and also they are seperated by the severity by color. As we can see only one point that is labeled fatal is clearly visible, meaniing that severity does not matter on the speed limit.

```
#no data wrangling needed here

ggplot(accidents, aes(x = longitude, y = latitude, color = severity, alpha = speed
  geom_point(size = 2)# graphing a geographic plot usin sccaterplot using longitud
```

```
## Warning: Removed 8 rows containing missing values or values outside the scale r
## (`geom_point()`).
```