Maths Skills 2 (Statistics) assessment

5 April 2018

Downloading the appropriate packages

Here I am loading in the required packages to perform the tasks.

```
library(tidyverse)
library(ggmap)
library(dplyr)
```

Loading the data

I created a new variable for each file then used the read_csv() function I read the data to that variable. I named each file according to the month in which those crimes were committed, So for example the variable "Jan" contains all the crime data from the January file. I used the the website https://data.police.uk/data/ to get all the crime data for my chosen city, Kingston Upon Thames.

```
Jan <- read_csv("~/2017-01-metropolitan-street.csv")
Feb <- read_csv("~/2017-02-metropolitan-street.csv")
Mar <- read_csv("~/2017-03-metropolitan-street.csv")
Apr <- read_csv("~/2017-04-metropolitan-street.csv")
May <- read_csv("~/2017-05-metropolitan-street.csv")
Jun <- read_csv("~/2017-06-metropolitan-street.csv")
Jul <- read_csv("~/2017-07-metropolitan-street.csv")
Aug <- read_csv("~/2017-08-metropolitan-street.csv")
Sep <- read_csv("~/2017-09-metropolitan-street.csv")
Oct <- read_csv("~/2017-10-metropolitan-street.csv")
Nov <- read_csv("~/2017-11-metropolitan-street.csv")
Dec <- read_csv("~/2017-12-metropolitan-street.csv")
```

Binding the data together

Using rbind() I created a new tibble "yearcrime" which is a combination of all the months from my data.

```
yearCrime <- rbind(Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec)
```

This would make it easier to make computations on all of my data in later stages.

Selecting chosen variables

I made another tibble "selectedCrime", selecting the only variables I needed; these were Month, Longitude, Latitude and Crime type.

```
selectedCrime <- yearCrime[c(2,5,6,10)]
```

The original data that I downloaded had 12 variables, doing this simplified it to 4.

Getting the coordinates for Kingston Upon Thames

Here I used the geocode() function to find the Longitude and Latitude coordinates for the centre of the city.

```
geocode("Kingston Upon Thames")
```

```
##   lon lat
## 1  NA  NA
```

These will be needed when using the get_map() function.

Map of Kingston Upon Thames

Below, I created a map of Kingston Upon Thames using get_map(). The location specifies Where I want the map. The source is the where I want to obtain the map from; here I chose Google Maps. The maptype is the type of map I want displayed. For example, I could have chosen "Satellite" which would have given me a satellite image of my chosen area.

```
mapKingston <- get_map(location = c(-0.300689, 51.41233), source = "google",
                       zoom = 14, maptype = "roadmap")
```

Creating a new variable mapKingston

I made a new variable "mapKingston" which contains my desired map. I did this so when I overlaid my data onto the map, I wouldn't have to repeatedly call the ggmap() function.

```
mapKingston <- ggmap(mapKingston)
```

Bounding Box

Here I created a bounding box containing the values in which my map is bounded by. I did this so I can filter out all the data which lies outside of Kingston Upon Thames.

```
bbox <- c(-0.35, 51.38, -0.25, 51.44)
```

Removing data that falls outside of my map

I used filter() to select the crimes which lie on the city map.

```
KingstonCrime <- filter(selectedCrime, bbox[1] <= Longitude, Longitude <= bbox[3],
                        bbox[2] <= Latitude, Latitude <= bbox[4])
```

Popular Crimes in Kingston Upon Thames

Here I used table() to show all the different types of crime that occurred. I did this so I could see which crimes occurred the most. These were Anti-social Behaviour, Violence and Sexual Offences, Burglary and Vehicle Crime. I needed this information so I knew which crime types to keep, and thus overlay onto my map.

```r
table(KingstonCrime$`Crime type`)
```

```
##
##        Anti-social behaviour                  Bicycle theft
##                         3374                            527
##                     Burglary      Criminal damage and arson
##                         1265                            916
##                        Drugs                    Other crime
##                          509                            134
##                  Other theft          Possession of weapons
##                         1226                             67
##                 Public order                        Robbery
##                          767                            183
##                  Shoplifting            Theft from the person
##                         1104                            336
##                Vehicle crime  Violence and sexual offences
##                         1329                           2965
```

Filtering the types of crime which were most common

I used filter() to select only the crimes which were most frequent. This bit of code is saying that if the data in my crime type column is equal to those characters, it would keep that piece of data.

```r
filteredCrimeKingston <- filter(KingstonCrime,
                         `Crime type` == "Anti-social behaviour" |
                         `Crime type` == "Violence and sexual offences" |
                         `Crime type` == "Burglary" |
                         `Crime type` == "Vehicle crime")
```

Checking to make sure that my new tibble only contained my selected crime.

This is checking that the only types of crime I have are infact the ones I selected.

```r
table(filteredCrimeKingston$`Crime type`)
```

```
##
##        Anti-social behaviour                       Burglary
##                         3374                           1265
##                Vehicle crime  Violence and sexual offences
##                         1329                           2965
```

Overlaying my data onto my map of Kingston Upon Thames

Here I mapped my data onto my map of Kingston Upon Thames using geom_point(). The aes() specifies how I want to put the data onto the map. I wanted to map the Longitude coordinates on the x-axis, the Latitude coordinates on the y-axis and I chose to change the colour of each crime so it's clear which crime types occurred. The size is the size of the dots, I chose the value 1 so it is clear exactly where the crime occurred. There was an option for shape, which changed the shape for each crime type, however this obscured the data as some shapes were hard to see and it wasn't clear where the crime occurred.

```
mapKingston + geom_point(aes(x = Longitude, y = Latitude, colour = `Crime type`),
                    size = 1, data = filteredCrimeKingston)
```
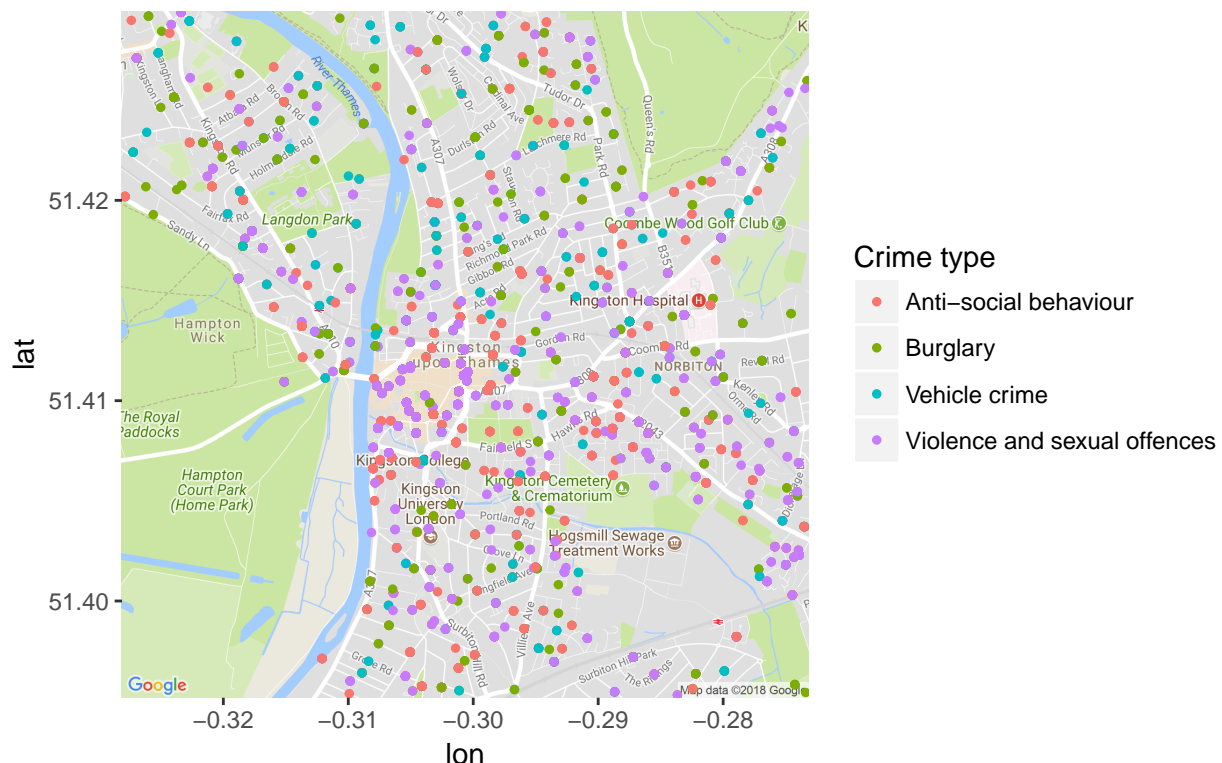

While this shows where the data lies on the map some of the data has been lost due to multiple crimes happening in the same location, thus each dot gets overlayed on one another. Furthermore, is hard to see if there is a area of Kingston which has a high amount of a certain crime. Another thing to comment on is that all the crimes occur in urban areas as oppose to the parks. This is to be expected as this is where the population density is highest.

Splitting my Crime data onto separate maps

Here I have done something similar to my first overlay. I added facet_wrap() to it which split up the graph into 4 sections, each section containing only one type of crime. I also had to decrease the size = 0.5 as the maps are now a quarter of the size and the dots weren't as accurate with size = 1.

```
mapKingston + geom_point(aes(x = Longitude, y = Latitude, colour = `Crime type`),
                         size = 0.5, data = filteredCrimeKingston) +
  facet_wrap(~ `Crime type`)
```


Here it is clear that there are far more crimes occuring in Kingston than previously observed as now there are fewer dots being layed onto themselves. It is now more evident which areas have more crime. For instance, there is a high density of Violence and Sexual offences in the city centre, particularly where there are nightclubs and bars.

Density of different Crimes

Below, I used geom_density2d() to map my data in another way.

```
mapKingston + geom_density2d(aes(x = Longitude, y = Latitude, colour = `Crime type`),
                             size = 0.5, data = filteredCrimeKingston) +
  facet_wrap(~ `Crime type`)
```
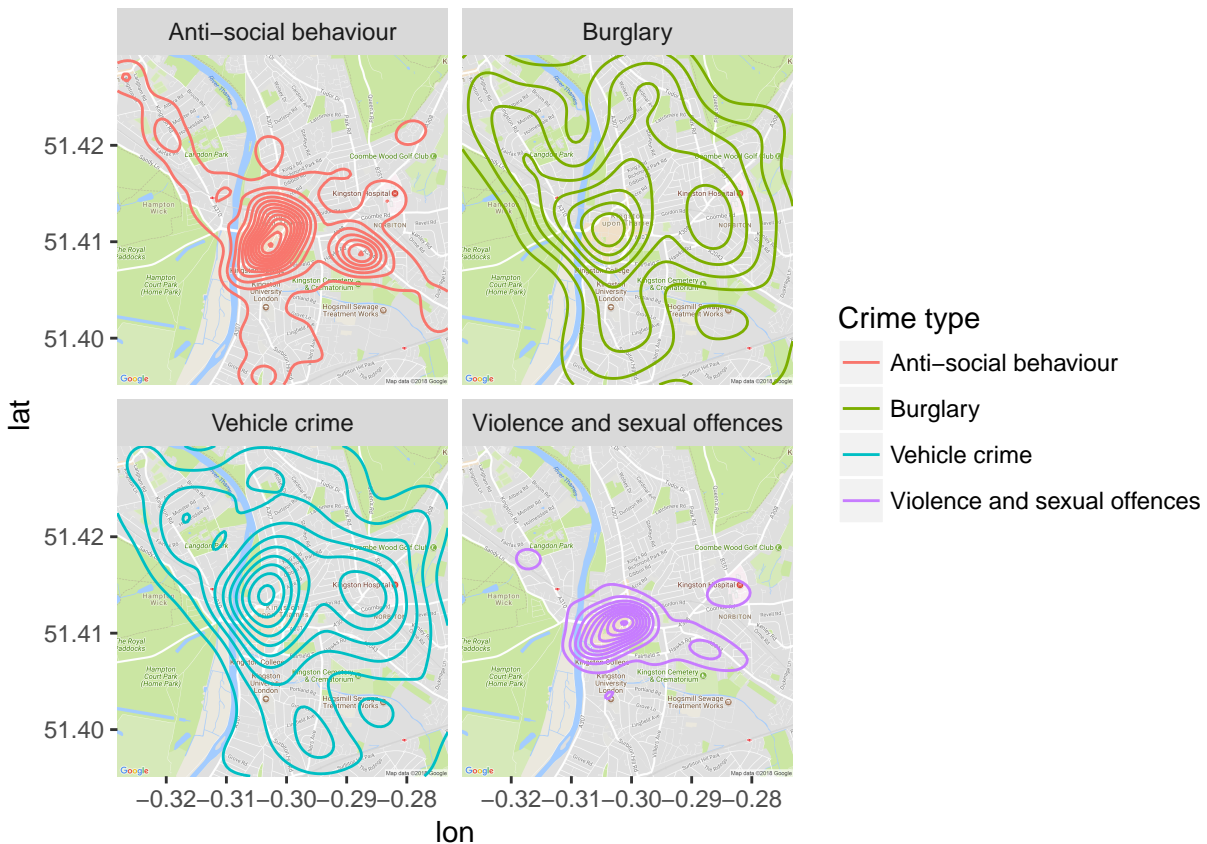

Here, each line represents a gradient of the frequency of that crime. This shows very clearly where the focus of a crime is. It is evident that Violence and sexual offences mainly occur in a very localised area. Moreover, it shows that a high amount of crime occurs in the centre. This could be due to many shops being in this area and a high density of people.

Barplot of all Crime in Kingson Upon Thames

Below, I filtered out unwanted columns. I only wanted the columns Month and Crime type.

```
kingstonSelectedCrime <- KingstonCrime[c(1,4)]
```

Converting Months into Seasons

Here I am changing the name for the Month. Firstly I change the data in the Month column so they can be read as characters. Then I change each month according to which season it lies in. So if the Month reads "2017-01", this is converted to "Winter". They are split like: Winter(Jan, Feb, Mar), Spring(Apr, May, Jun), Summer(Jul, Aug, Sep), Autumn(Oct, Nov, Dec).

```
kingstonSelectedCrime$Month <- as.character(kingstonSelectedCrime$Month)
kingstonSelectedCrime$Month[kingstonSelectedCrime$Month == "2017-01" |
                            kingstonSelectedCrime$Month == "2017-02" |
                            kingstonSelectedCrime$Month == "2017-03"] <- "Winter"
kingstonSelectedCrime$Month[kingstonSelectedCrime$Month == "2017-04" |
                            kingstonSelectedCrime$Month == "2017-05" |
                            kingstonSelectedCrime$Month == "2017-06"] <- "Spring"
kingstonSelectedCrime$Month[kingstonSelectedCrime$Month == "2017-07" |
                            kingstonSelectedCrime$Month == "2017-08" |
                            kingstonSelectedCrime$Month == "2017-09"] <- "Summer"
kingstonSelectedCrime$Month[kingstonSelectedCrime$Month == "2017-10" |
                            kingstonSelectedCrime$Month == "2017-11" |
                            kingstonSelectedCrime$Month == "2017-12"] <- "Autumn"
```

Creating a Factor Variable

Here I am creating a new tibble which is a factor variable "factor_kingstonSelectedCrime". The factor Variable has levels This will allow me to use the data to create plots.

```
factor_kingstonSelectedCrime <-
  mutate(kingstonSelectedCrime, `Crime type` =
           factor(`Crime type`, levels = unique(`Crime type`)))
```

Amount of crime that occurs across each season

Using ggplot() I made a stacked barplot. This shows the frequency of different types of crime. Each bar is split into the season in which the crime occured. Firstly, I selected the data which was the new factor variable. Then in aes() I specified what I wanted displayed on the x axis as well as the fill. The fill caused the barplot to be split into each season. Lastly, I used theme() to rotate the labels as they were overlapping.

```
ggplot(data = factor_kingstonSelectedCrime) +
  geom_bar(mapping = aes(x = `Crime type`, fill = Month) ) +
  theme(axis.text.x = element_text(angle = 90))
```
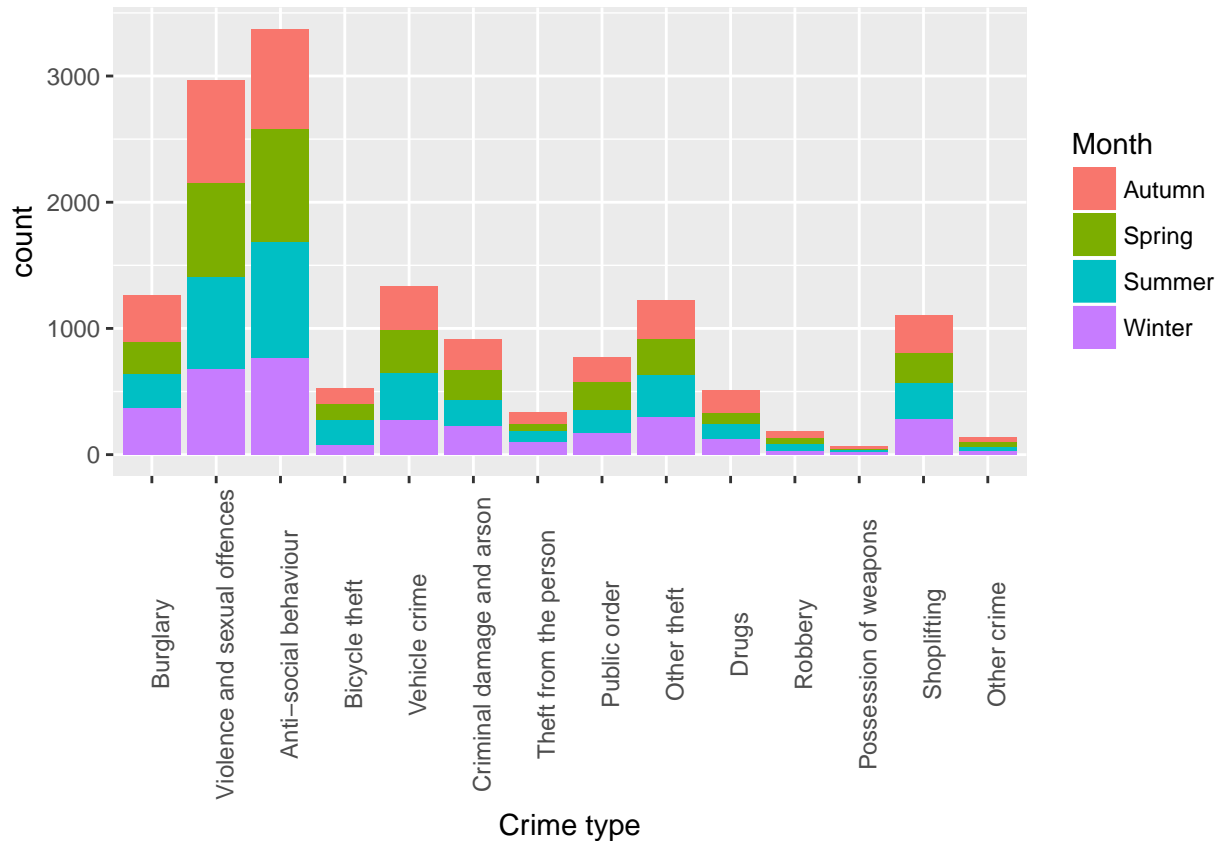

This shows the frequency for different types of crime and each bar is split into the season in which the crime occured. For each crime the bar is split evenly, this suggests that a particular season does not attracts more crime. However, for crimes which happen less frequently, such as "Possesion of weapons", it is hard to distinguish whether a season is subject to more crime. If we wanted to investigate whether a certain crime occured for a particular season, we would need to compare this data with other years.

Distribution across Seasons of different Crime

Here I did something similar to the above, however I added "position ="fill"". This created a distribution so you can see the percentage of what crime happened for all the seasons.

```
ggplot(data = factor_kingstonSelectedCrime) +
  geom_bar(mapping = aes(x = `Crime type`, fill = Month), position = "fill") +
  theme(axis.text.x = element_text(angle = 90))
```
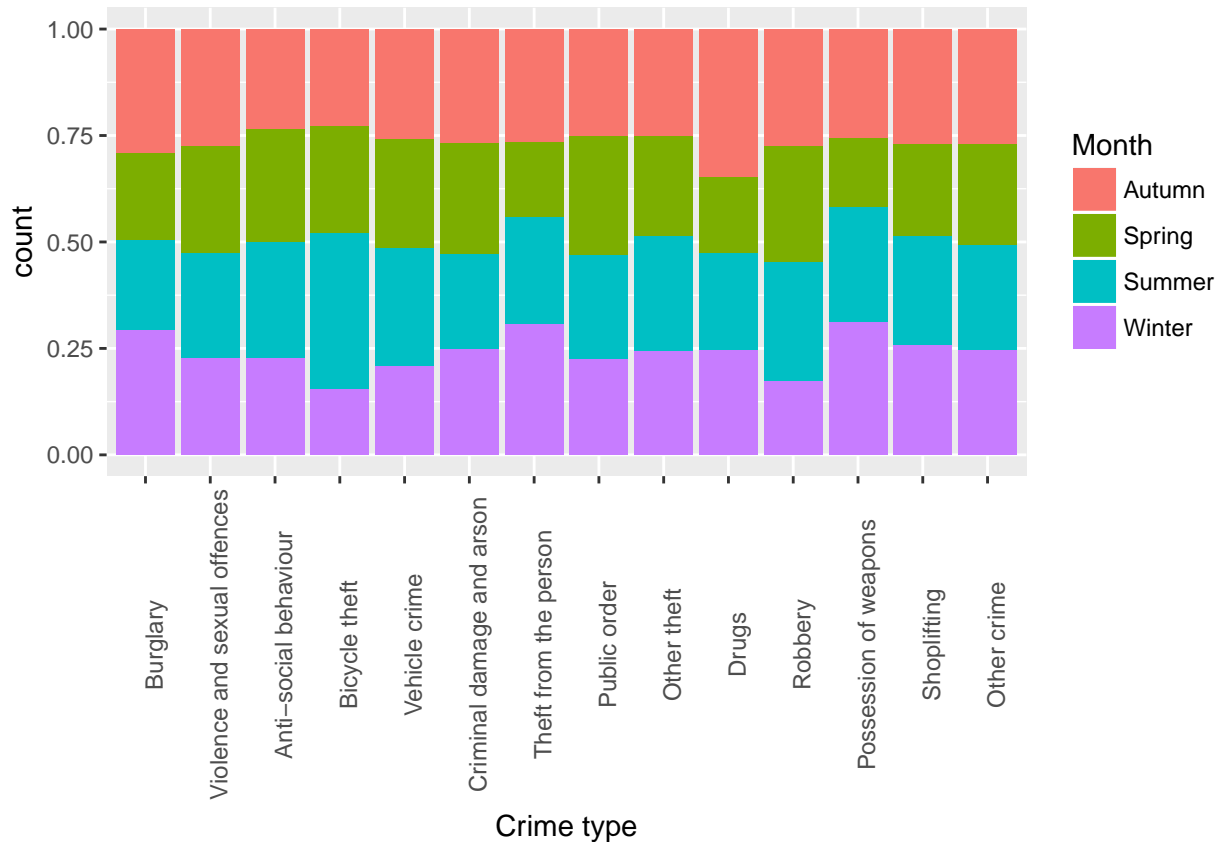

This is a distribution of how much of one crime occured in all the seasons. This further support that a particular season does not attract more crime. Moreover, for crimes that occur less frequently, it is likely that the variance of the height of each bar is going to be greater. One would have to collect more data to accurately interpret whether a particular crime was characteristic of a certain season.