Question 1

i) The dataset has only two variables:

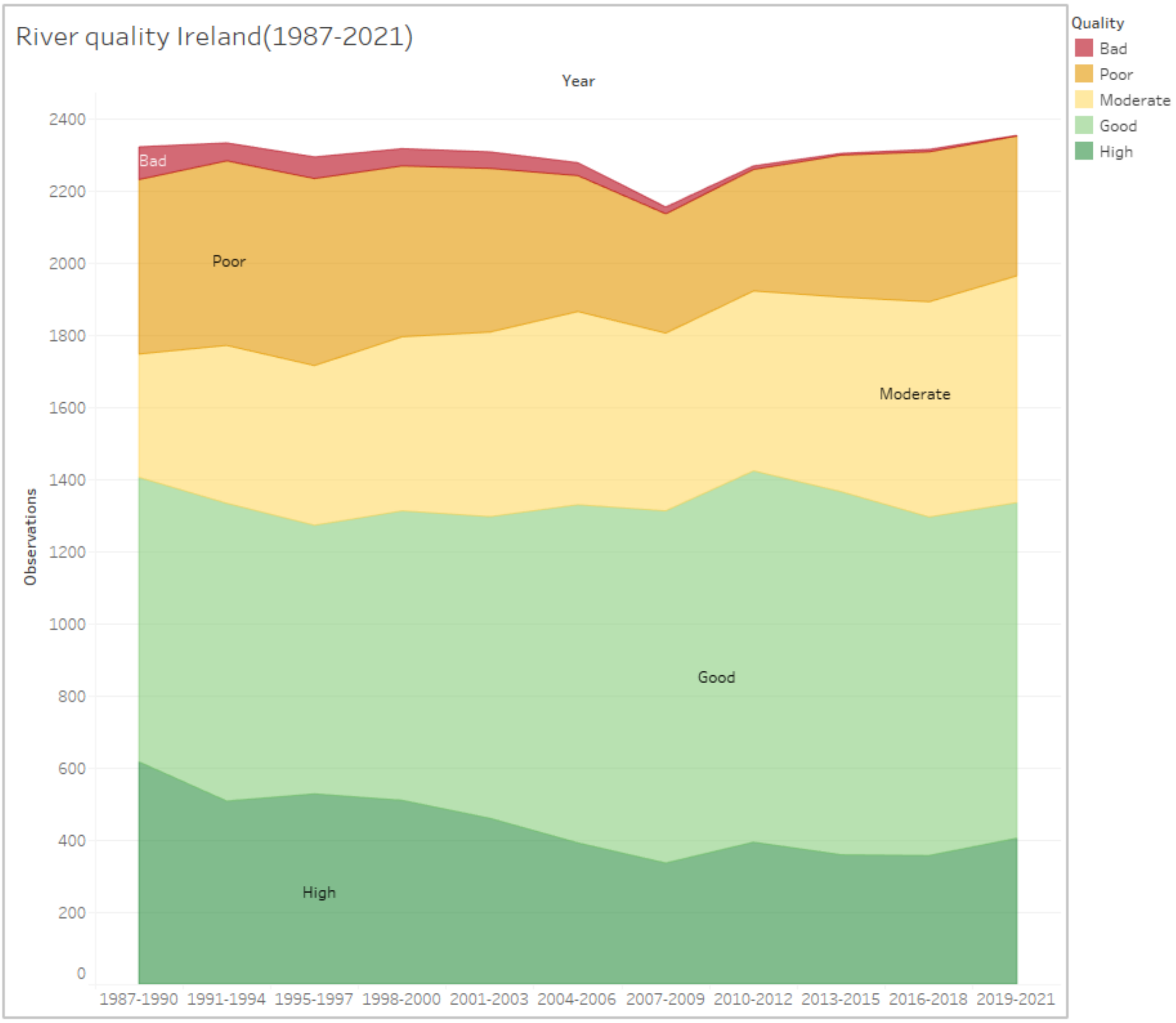
Year: the time period in which the observations have been gathered (continuous)

Year is a continuous variable (even though is presented in batches in this case) as it can be measured quantitatively and it's not limited by a specific category

Quality: an ordinal qualitative variable which describe the state of the river at the time of the observation (discrete)

The main feature of discrete variables such as Quality of water is that the values cannot be measured but can indeed be counted, also the possible values are limited by a specific category: in this case as there are only 5 possible values for quality.

ii)



iii) I believe the main problems in the way data is shown in this plot can be categorised in three sections:

1. Direction: the plot represents time vertically on the y axis in a descent motion (recent to older), it would have been better to locate the time variable on the x axis acting as temporal line or at least going with an inverse order (older to recent) to help the reader following the timeline while reading from top to bottom.

2. Visual impact: the bars show every single data point in terms of number of observations while the overall plot conveys data as a percentage, on one hand this is great because it’s showing a lot of data but in the eye of the average reader of the newspaper it might result in a confusing picture with all this numbers; the same is especially true for showing the total number of observations on each year which, in my opinion, does not bring any value to the average reader of the Irish Times; in summary, I would have preferred making more by showing less.

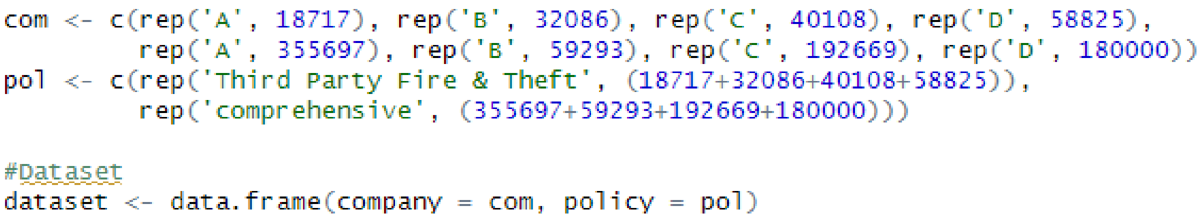
3. Legend location: it would have been more appropriate to post the legend on top so that a person reading through would know the meaning of the colours right from the beginning.

Question 2

i) There are only two variables:
- Company
- Type of Policy

ii) both variables are discrete, for both variables the possible values are restricted by a category, for both of them it is not possible to quantitatively measure a value against another but it is possible to count their occurrences.

iii) The dataset has been transformed into a long dataset in R by this line of code:



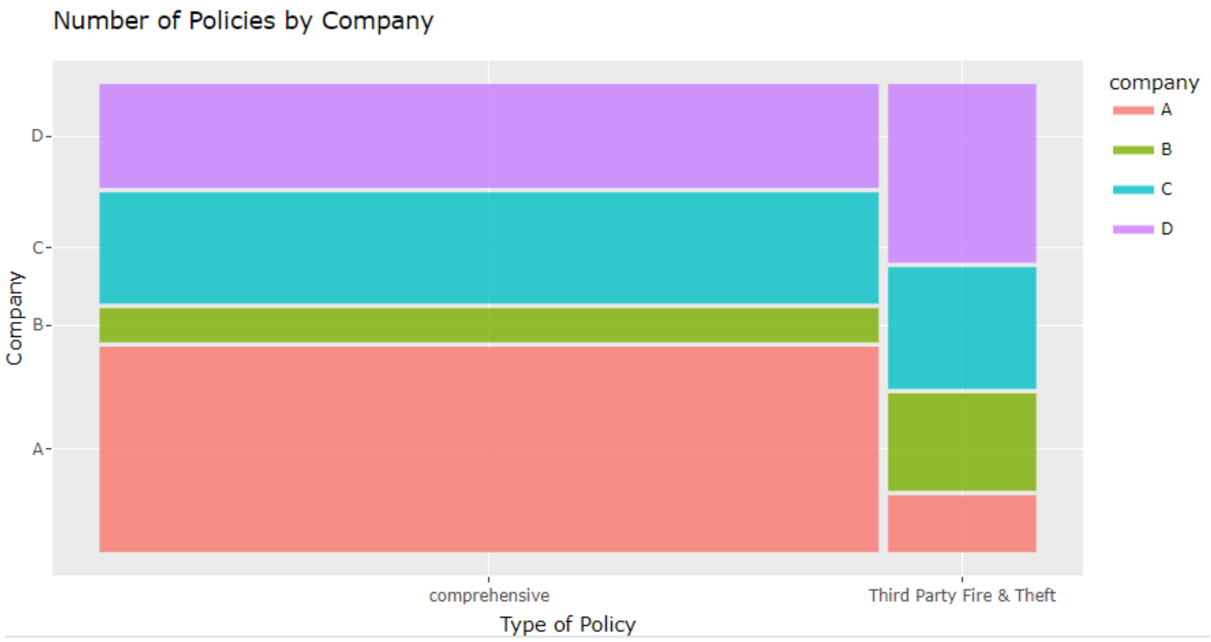
```
com <- c(rep('A', 18717), rep('B', 32086), rep('C', 40108), rep('D', 58825),
         rep('A', 355697), rep('B', 59293), rep('C', 192669), rep('D', 180000))
pol <- c(rep('Third Party Fire & Theft', (18717+32086+40108+58825)),
         rep('comprehensive', (355697+59293+192669+180000)))

#Dataset
dataset <- data.frame(company = com, policy = pol)
```

In the two variables (com and pol) the values for company name and policy type have been repeatedly inserted for the amount shown in table, the two variables then have been combined into a dataframe

iv) Using graph paper draw a mosaic plot by hand using graph paper. Take an image of the plot, all of your calculations and paste them into your report.

v)



vi) it is obvious that the distribution of policies provided by company appear to be very different for the two selected policies, almost inverted with company A being the leading force in comprehensive policies while having the least number of policies for third party and theft policies; company D on the other hand performs very well in the latter compared to comprehensive policies.
If we were to rank the companies by market shares based on policy we would see two completely different orders as:

Highest to lower for theft and fire: D,C,B,A
Highest to lower for comprehensive: A,C,D,B

vii) there might be multiple reasons for the relationships between insurance company and type of policies:

1. Location and market area of the company: it might be that the companies are located in areas with very different theft instances.
2. Discounts: it might easily be that some company offers different discount rates for some specific policy rather than for others therefore resulting more appealing to the customers.
3. Market sectorization: it is possible that a company prefers to specialise in a specific policy therefore granting higher shares for that specific market.

Question 3

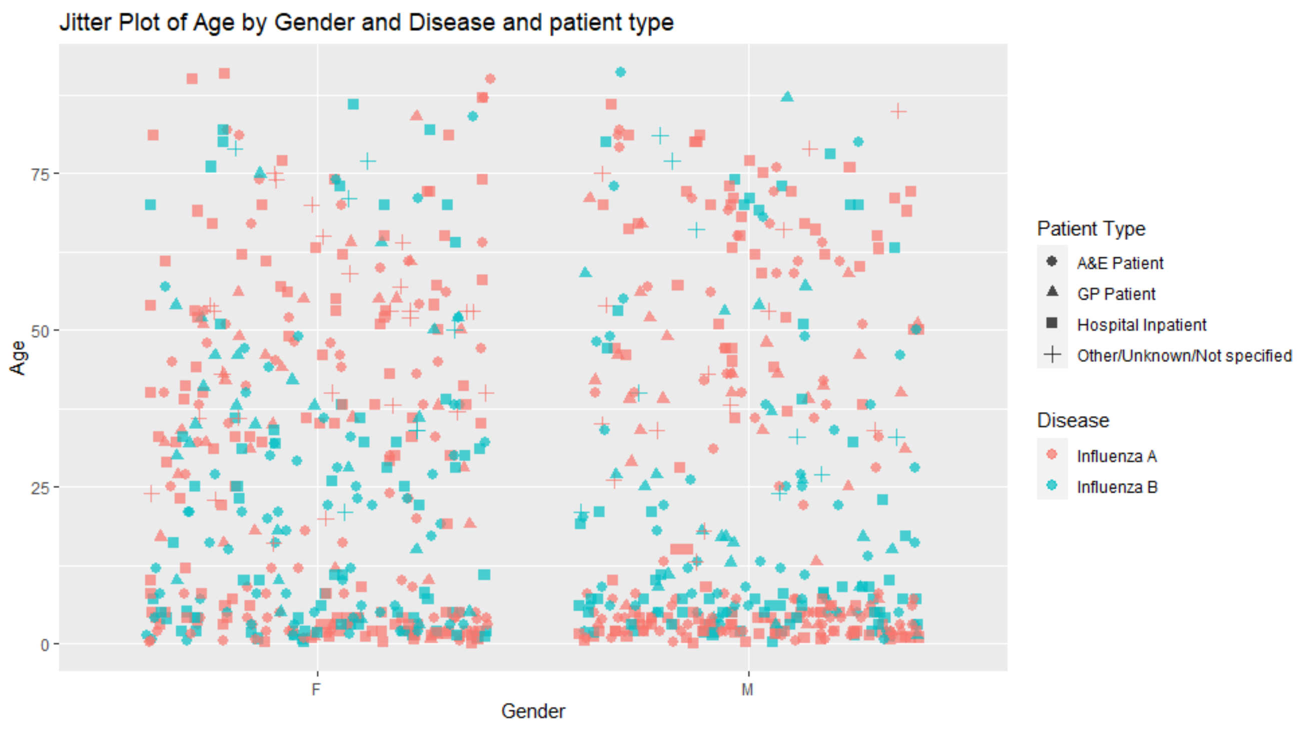
i) the variables are: Patience type (discrete), Disease (discrete), Gender (discrete), Age (Continuous)

Ii - iii) All 4 files are in the submission as html files

iv) Influenza in this region is very common, however not all influenza is the same; influenza A is characterised by being extremely common among kids of both sex; however this type of influenza is more common among females than males especially for young adults.
Influenza B on the other hand showcases similar characteristics for both males and females of all age groups, also for this type kids appear to be more impacted than adults.

v) I would use the histogram trellis plot as the data shown does a good job in showcasing the age distribution in a simple way by simply showing four histogram distributions for influenza type and gender with not too many bins; main drawback of the plot is that it doesn’t do a good job in showing which type of patience is more affected but since this plot is meant for general use the variable patience type is not that important compared to gender, influenza type and age.

vi) Another possible graphic for visualising all four variables might be a jitter plot with age and gender on the two main axis as they might be the most relevant variables, followed by the type of influenza expressed as a colour and lastly the type of patient as the shape of the point.
This graphic can be great for visualising possible clusters of patients.



Question 4

a-b) The file in submission is called CA.twbx; the dashboard contains all the four worksheets and by using actions it creates an interactive environment where by clicking on the disease name in the age distribution it shows the incidence over time for that disease only and correspondent location of patients, same applies when clicking on the gender distribution.

c) The dashboard is extremely useful for comparing the three diseases:

- Cryptosporidiosis: it looks to be more impactful among young patients as the median age for this disease is 10 years old; it is also more prevalent in women by a substantial margin and after firsts spikes in 2012 the number of patients seems to have increased after 2015.

- Giardiasis: with a median age of patients as 31 years old it seems to be less impactful in kids than Cryptosporidiosis even though it is still present among young patients; it is also more prevalent in men by a substantial margin and after the first spikes in 2014 the number of patients seems to have increased dramatically following 2015.

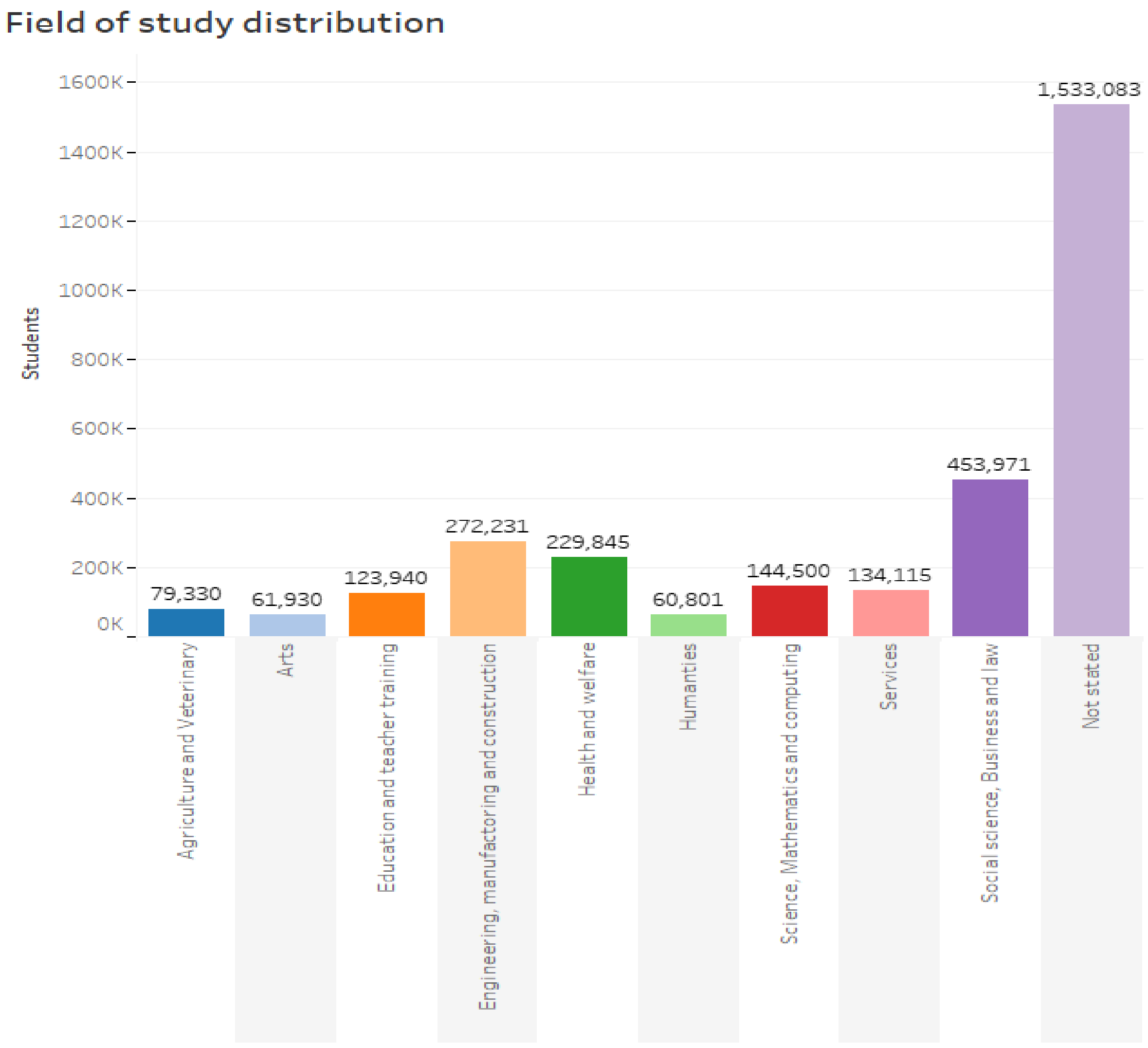
- Verotoxigenic Escherichia: definitely the most impactful overall with the highest number of patients from the dataset, median age of patients is 22, more present in women than men even thought by a lower margin compared to Cryptosporidiosis; in terms of incidence over time it looks having a higher variation compared to the other two but again spikes seem to have increased after 2015.

Overall the three diseases together appear to have a particular geographic distribution, South Dublin appear to have higher numbers of patients, especially for Verotoxigenic Escherichia and Cryptosporidiosis, the geographical highlights a probable outbreak of Verotoxigenic Escherichia in Lucan.

Question 5

I to iv) the dashboard is saved as “question5” in the file CA.twbx

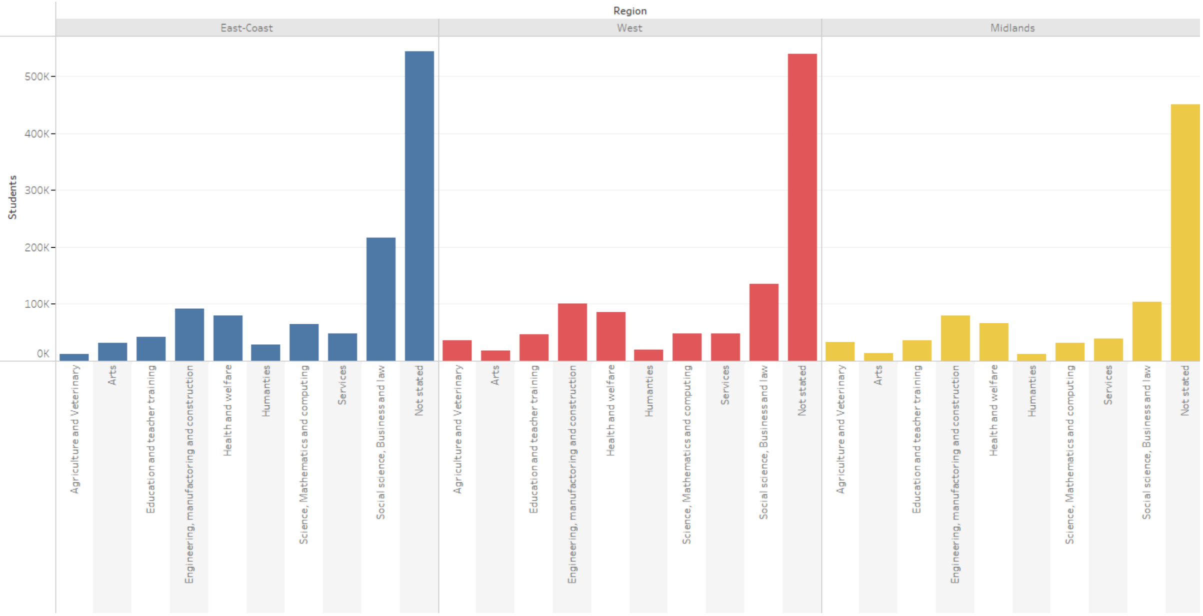
vi) the variable chosen was the field of study reported for student of 15 years old or older; the overall distribution for the whole country is summarised in this chart:



Field of study distribution

As shown, most of the people did not report the selected field of study, probably because the student in the household didn’t choose yet a specific field; however overall social studies Business and Law seems to be the most popular field followed by Health and engineering, Arts and humanities appear to be the least desired among students with the lower rates.

Field of study distribution



A closer look to the three regions shows that the overall distribution of field of study is quiet similar among the regions but with some important differences, for example the percentage of students pursuing an art related degree is higher in the East-coast region compared to the rest, business and law is the most popular field of study across all the regions but with a greater margin for the East-coast region; engineering is quite popular also across all regions while agriculture appear more popular in midlands and West compared to the East-coast, arts appear to be the least popular in the West region while humanities in the midlands.

b)Maps serve as powerful tools for visualizing progress towards the Sustainable Development for multiple reasons. maps provide a spatial context that enables a comprehensive understanding of the distribution and concentration of various indicators across regions. this is crucial for identifying disparities, potential clusters and areas of interest.

Maps can also facilitate the identification of trends and patterns. By simply overlaying and comparing maps for different themes (for example socioeconomic or demographic maps) it becomes easier to discern correlations and causal relationships.

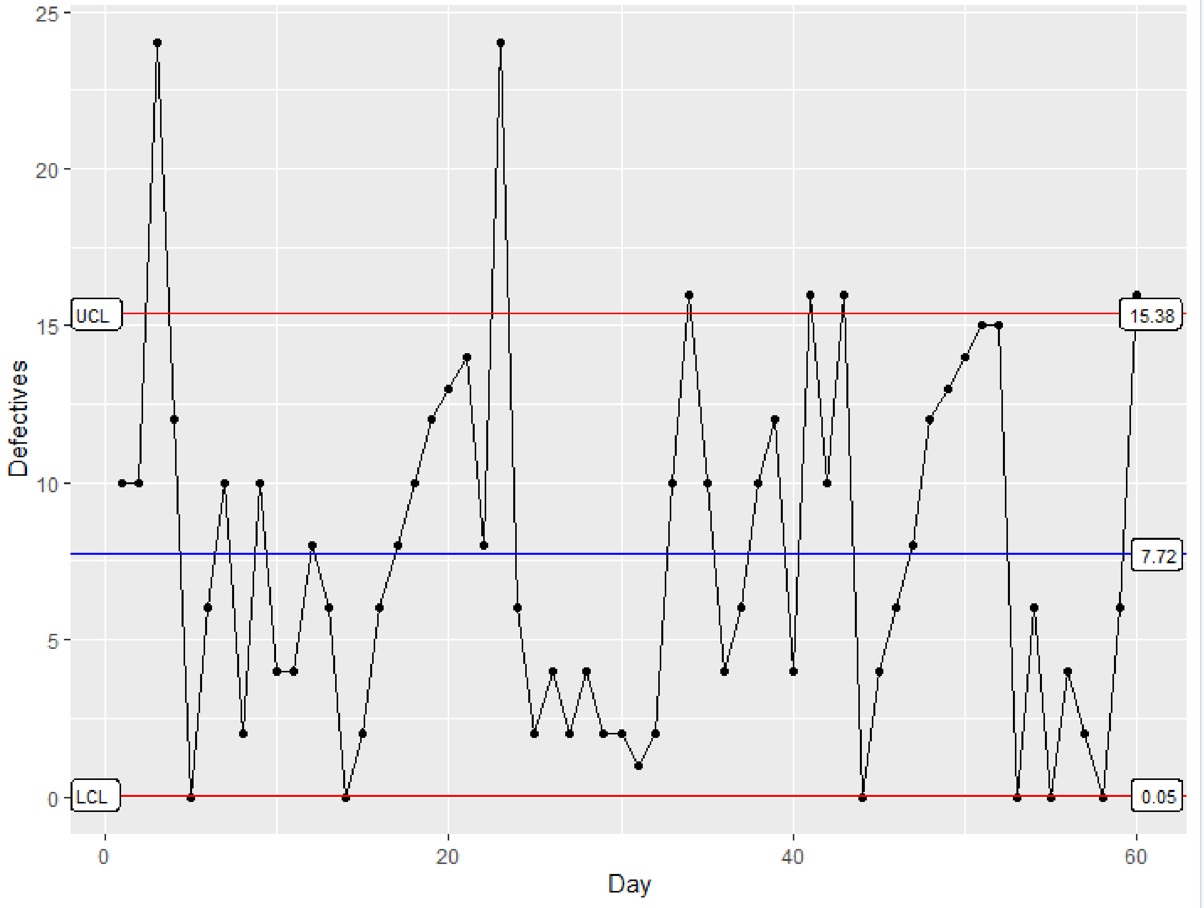
The same is true for overlaying maps for different periods for assessing the impact of policies and interventions, as well as for predicting future trends.

Maps also offer a visual language that is accessible to a broad audience, including policymakers, communities, and the general public. Complex data sets related to the SDGs can be translated into clear visuals, developing a shared understanding of the progress made and possible challenges. This accessibility promotes transparency and accountability, encouraging collaboration among stakeholders.

In summary, maps provide a holistic and dynamic representation of progress towards the SDGs, incorporating spatial, temporal, and visual dimensions. They are indispensable tools for policymakers, researchers, and communities alike, facilitating informed decision-making.

Question 6

i)



ii) P is calculated as:

Total number rejected= 463

Total Number inspected = 60 days * 50 per batch = 3000

$P = 463/3000 =0.15$

$nP = 0.15*100 = 15$

$CL = 3\sqrt{nP(1-P)}= 3\sqrt{15(1-0.15)} = 3\sqrt{12.75} = 3*3.5 = 10.5$

$UCL = nP + 10.5 = 15.4 + 10.5 = 25.9$

$LCL = nP - 10.5 = 15.4 - 10.5 = 5.4$

Iii - iv) Out of 60 days, the SPC chart showed that multiple times the process have reached out of control points, also in two separate instances (around the 15th and 45th day) the plot shown a positive trend: this might be an indicator of changing factors in the process as the main assumption of an SPC control plot is that the results should follow a random variation and not positive or negative trends

v) In Statistical Process Control (SPC); charts, such as the one you've been shown, the concept of control limits is fundamental to monitoring and ensuring the stability of a process. Control limits are statistical boundaries that define the expected variation in a process when only common causes of variation are present. Common causes are inherent to the process and include factors like machine variation, raw material differences, or changes in environmental conditions. The upper and lower control limits are typically set at a certain number of standard deviations from the process mean. In a stable process operating under normal conditions, nearly all data points should fall within these limits. The purpose of these limits is to distinguish between normal, expected variation (within the limits) and exceptional, unexpected variation (outside the limits), which may indicate the presence of special causes. By monitoring the process within these statistical bounds, the SPC chart helps identify when the process is operating as expected and when there might be a need for investigation and corrective action due to unusual variability or shifts in the process.

b) A significant decrease in failure numbers could indicate a shift in the process mean, which may require investigation. While a lower failure rate is generally positive, a shift might suggest a change in the process that needs attention.

c) Another possible use for SPC chart can be in testing the overall satisfaction of a customer where a good amount of indicators might come together and define a satisfaction score between 0 and 10, at that point using historical data a SPC chart can be drawn and from that time on customer scores can be put into the SPC chart to control if changes in satisfaction levels are randomly generated or not.