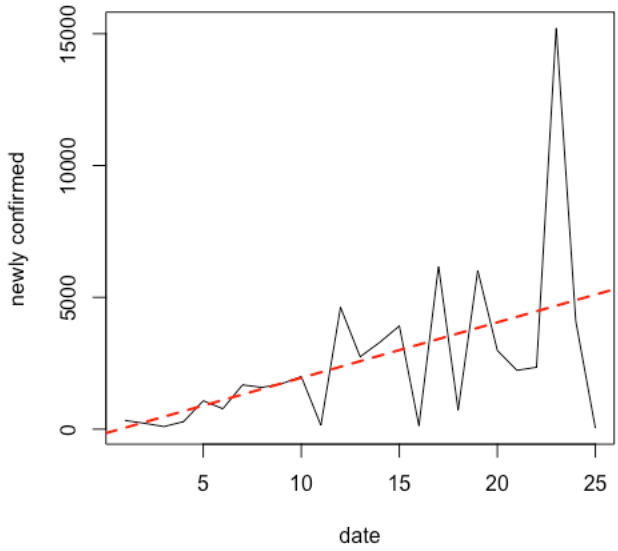
Name: CHAN King Yeung
SID: 1155119394
SEEM2460 Assignment 2

Question 1
The novel coronavirus has made a huge impaction around the world. The number of confirmed cases is growing in an uncontrollable state. As safety is the priority in such circumstance, classes are suspended, and many companies recommend their employees to work from home. Data visualisation gives us an insight into the behaviour of data, such that we can make a quick decision. However, every data visualisation has different usages, which depends on the purpose of research. Although heat map does a great job in finding the severest region, being a nerd in the current situation, the line plot does address the concern of whether the virus has controlled or not. From the given data, we have the following line plot illustrating the number of newly confirmed cases per day.



The x-axis refers to the date from 21 January to 14 February, whereas the y-axis indicates the number of newly confirmed cases. The red dashed line in the graph is the regression line for fitting the data. The regression line shows that the number of newly confirmed cases is still increasing at the moment. It is crazy that the highest point in the graph is about 15,000 newly confirmed cases in one day. After retrieving from the original data, there were 15,211 newly confirmed cases on 12 February, and what more important is, 99.82% are from Mainland China.

I currently cannot imagine more than two cases that make such happen. The first one is the original data sets contain errors. The other one, which I think is likely to be, it does happen in Mainland Chain. (my opinion omitted here since it is out of topic). If the second scenario is true, thanks to China, the graph tells us that the virus has not controlled yet, and the number of newly confirmed cases tends to increases at the moment.

Question 2

a) Let X be the number of text messages sent per day and Y be the average hours of sleep

$\bar{x} = \frac{\sum X_i}{n} = \frac{585}{5} = 117$ text messages

$s_X = \sqrt{\frac{n\sum X_i^2 - \sum X_i^{\,2}}{n(n-1)}} = \sqrt{\frac{5(82963)-(585)^2}{5(4)}} \approx 60.2453$ text messages

$\bar{y} = \frac{\sum Y_i}{n} = \frac{26.6}{5} = 5.32$ hours

$s_Y = \sqrt{\frac{n\sum Y_i^2 - \sum Y_i^{\,2}}{n(n-1)}} = \sqrt{\frac{5(154.92)-(26.6)^2}{5(4)}} \approx 1.8308$ hours

b) $SXX = \sum x^2 - n\bar{x}^2 = (82963) - 5(117)^2 = 14518$, $SYY = \sum y^2 - n\bar{y}^2 = (154.92) - 5(5.32)^2 = 13.408$,
$SXY = \sum xy - n\bar{x}\bar{y} = (2699.2) - 5(117)(5.32) = -413$

$r = \frac{SXY}{\sqrt{SXX \cdot SYY}} = \frac{-413}{\sqrt{(14518)(13.408)}} \approx -0.9361$

As r is close to -0.9361, there is a strong negative association between the average hours of sleep and text message per day

---

Question 3

a)
```
SELECT StudentName
FROM Student, Course, CourseEnrollment
WHERE Student.StudentID = CourseEnrollment.StudentID AND
        CourseEnrollment.CourseID = Course.CourseID AND
        Student.Department = 'Math' AND
        Course.CourseID = '1001'
```

b) $\prod_{StudentName}\left(Student \bowtie \left((\sigma_{CourseName='Calculus'}Course) \bowtie CourseEnrollment\right)\right)$

Question Bonus

One trivial way to understand the reason for dividing by $n-1$ is that it is the unbiased estimator for the population standard deviation. There are altogether n degrees of freedom for a sample of size n. After drawing the sample of size n, all the value of x's are known, hence there is an extra equation $Y \coloneqq \sum X_i = n\bar{X}$, which create a restriction from the sample. Since the sample variance uses the sample mean to derive, given the fact that they are independent, there is 1 degree of freedom lost in the distribution of sample variance $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$. Therefore, the degrees of freedom for the sample standard deviation is now $n-1$.