36-402 Homework 1

Eu Jing Chua
eujingc

January 20, 2019

Question 1

Table 1: Summary of of data

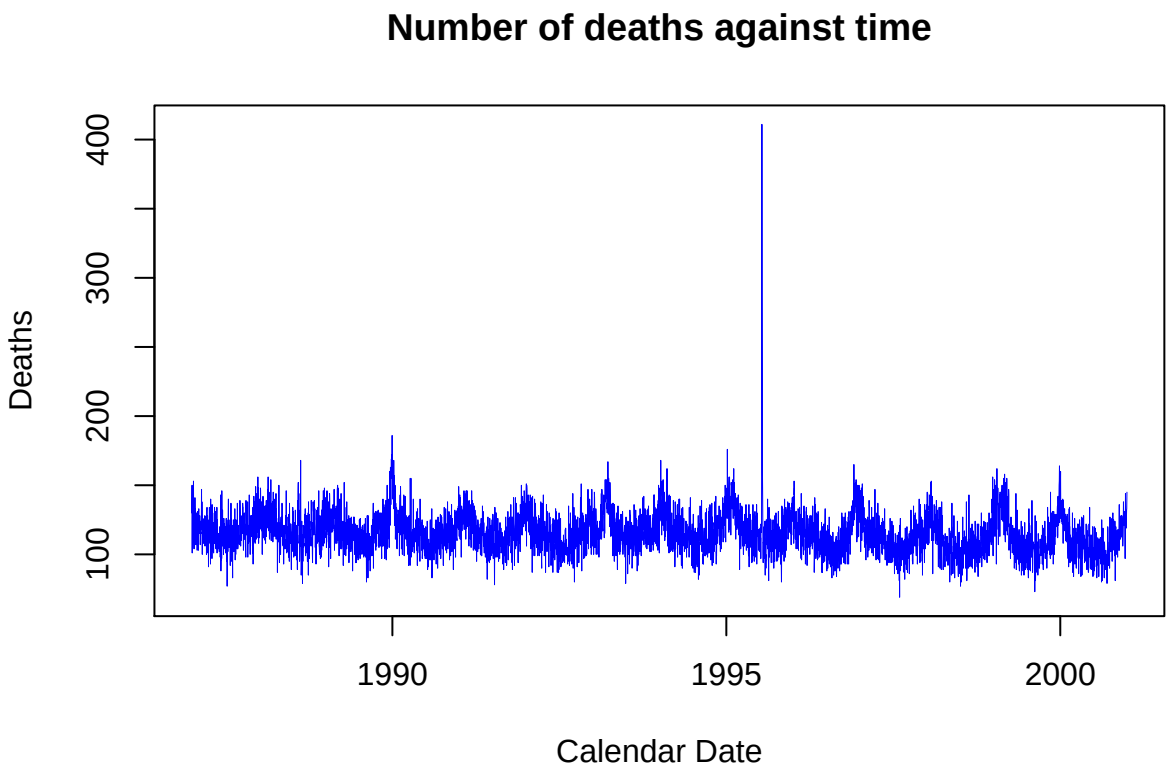
death	pm10median	pm25median	o3median	so2median	time	tmpd
Min. : 69	Min. :-37.376	Min. :-16.43	Min. :-24.78	Min. :-8.206	Min. :-2556	Min. :-16.0
1st Qu.:105	1st Qu.:-13.108	1st Qu.: -6.59	1st Qu.:-10.23	1st Qu.:-2.689	1st Qu.:-1278	1st Qu.: 35.0
Median :114	Median : -3.539	Median : -1.33	Median : -3.33	Median :-1.218	Median : 0	Median : 51.0
Mean :115	Mean : -0.146	Mean : 0.24	Mean : -2.18	Mean :-0.636	Mean : 0	Mean : 50.2
3rd Qu.:124	3rd Qu.: 8.303	3rd Qu.: 5.34	3rd Qu.: 4.47	3rd Qu.: 0.832	3rd Qu.: 1278	3rd Qu.: 67.0
Max. :411	Max. :320.725	Max. : 38.15	Max. : 43.69	Max. :28.903	Max. : 2556	Max. : 92.0
NA	NA's :251	NA's :4387	NA	NA's :27	NA	NA

1 a) The temperature is given in Fahrenheit.

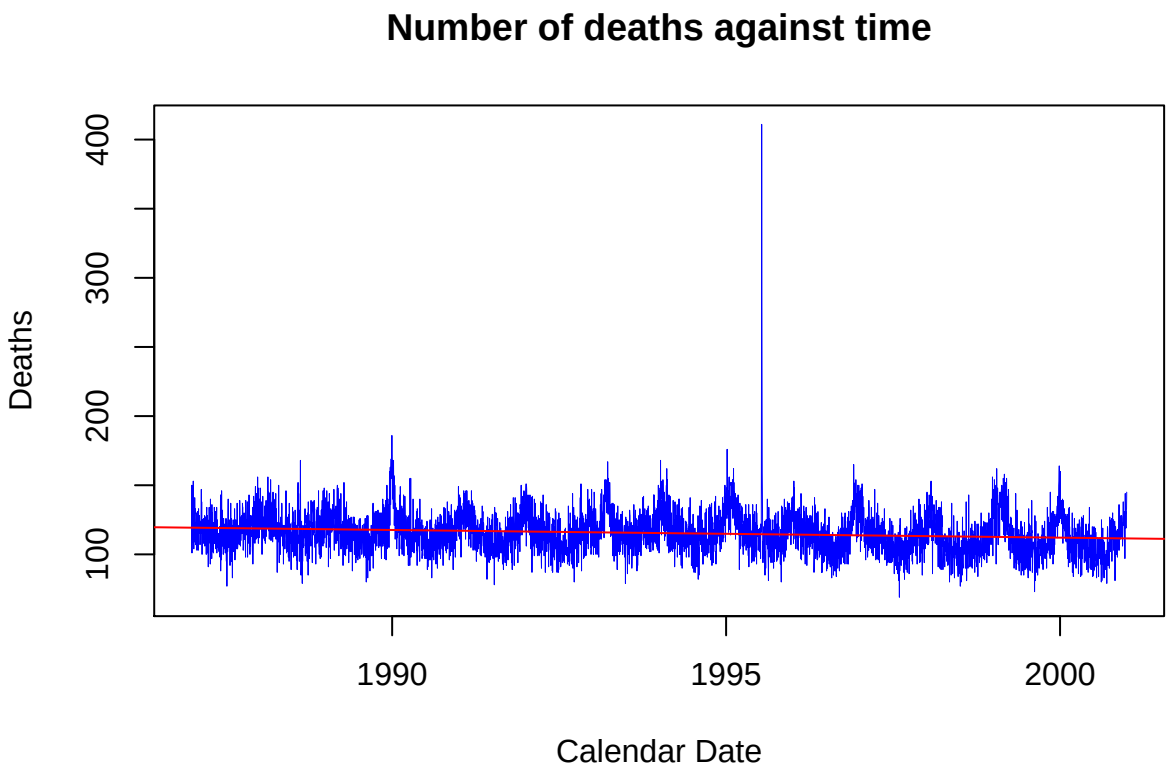
1 b) This indicates the data has probably been log-transformed as it does not make sense to have a negative density of particles.

Question 2

Q2 a)



There seems to be a seasonal, roughly sinusoidal pattern to the number of deaths across time, where the cycle period is approximately a year. There is also a sharp peak between 1995 and 1996. The code fragment transforms the `time` offset from the first measurement into calendar dates, starting from 31 Dec 1993 and then stores these dates into a new column `date`.



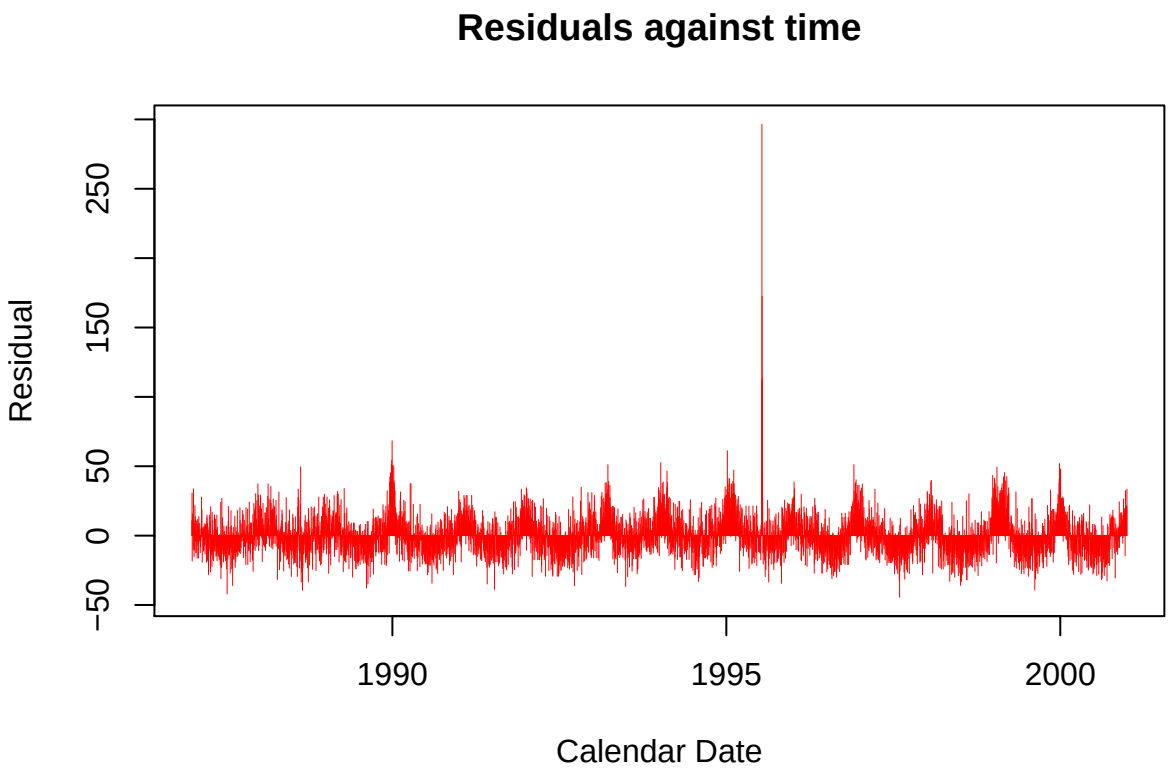
Q2 b)

Table 2: Coefficients & p-value

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	128.7478	1.2742	101.0399	0
date	-0.0015	0.0001	-10.6077	0

As the p-value of the slope is very close to 0, we can conclude that the slope is significantly different from 0, since it has such a low standard error.

Q2 c)



There is a similar seasonal, sinusoisal pattern in the residuals with a similar sharp peak between 1995 and 1996, instead of the expected random scatter around 0 with constant variance. Hence, these residuals exhibit correlation across.

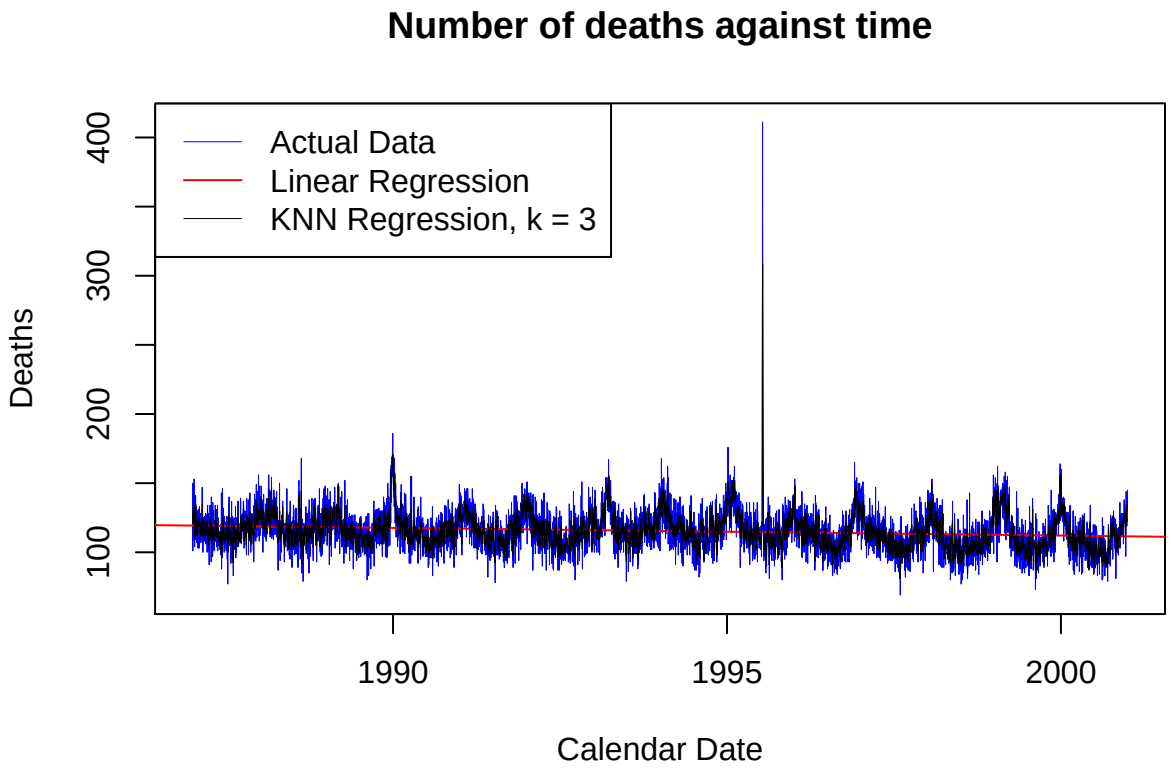
Q2 d)

Given that we assume the assumptions of linear regression by ordinary least squares hold, the regression slope indicates that for each passing day, we predict the number of deaths decreases by 0.0015 on average.

Q2 e)

We have reason to doubt the validity of the significance test here as the residuals are in major conflict with the initial assumptions of the linear regression we have performed. It is unreasonable to think the true trend of the data as-is is actually linear, hence we have reason to doubt the validity of the significance test of the slope.

Question 3

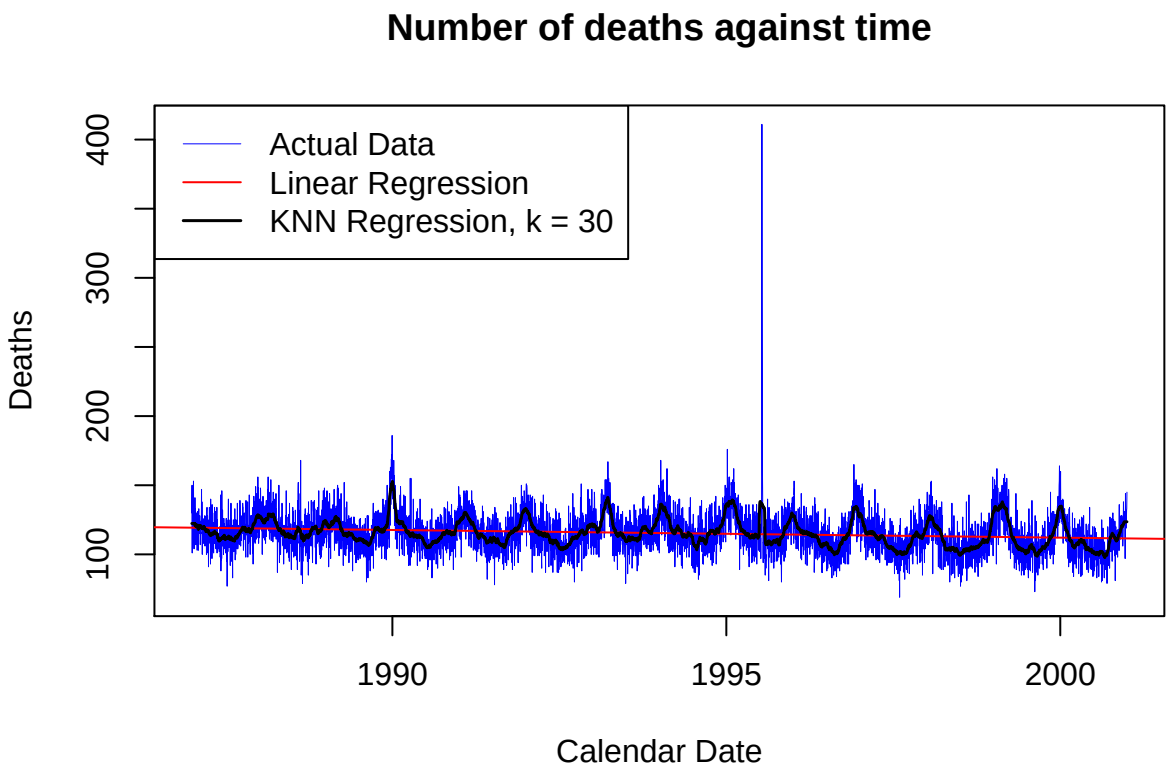


Q3 a)

The shape of the estimated function closely resembles that of the original function but slightly less noisy. It is periodic and sinusoidal, with the same sharp peak between 1995 and 1996.

Q3 b) For each point in time t, its predicted value $\hat{x}(t)$ is derived from the simple average of the 3 closest points in time from the original data, i.e. $\hat{x(t)} = \frac{1}{3}(x(t-1) + x(t) + x(t+1))$, since we have regular intervals of time, which is similar to a moving average with sliding window of size 3.

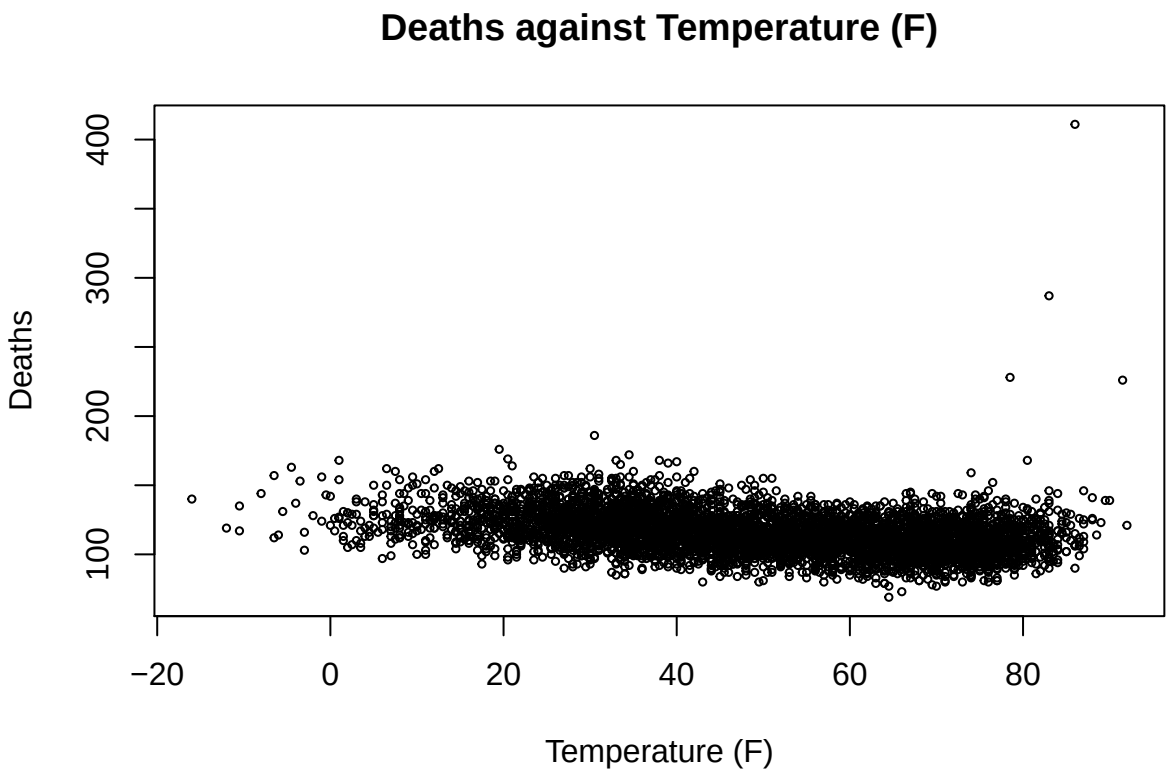
Q3 c)



This new estimate of the curve is much less noisy than before, being much smoother. It still exhibits the same overall sinusoisal pattern, but with a much less dramatic spike between 1995 and 1996.

Question 4

Q4 a)



There seems to be a rough negative linear trend in the plot above, with some signs of homoscedasticity throughout. However, there also seem to be several outliers from this trend past 80F, where the number of deaths are much higher. There is also many more data points between 20F to 80F as compared to outside this range.

Q4 b)

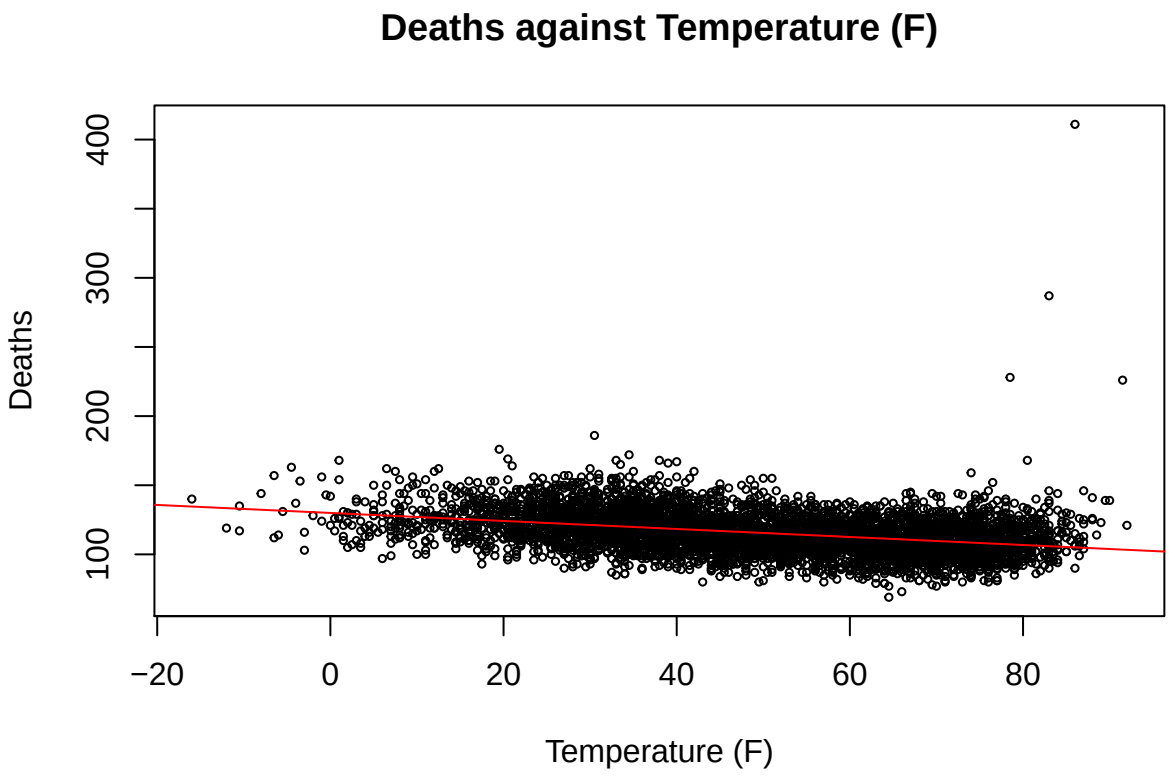


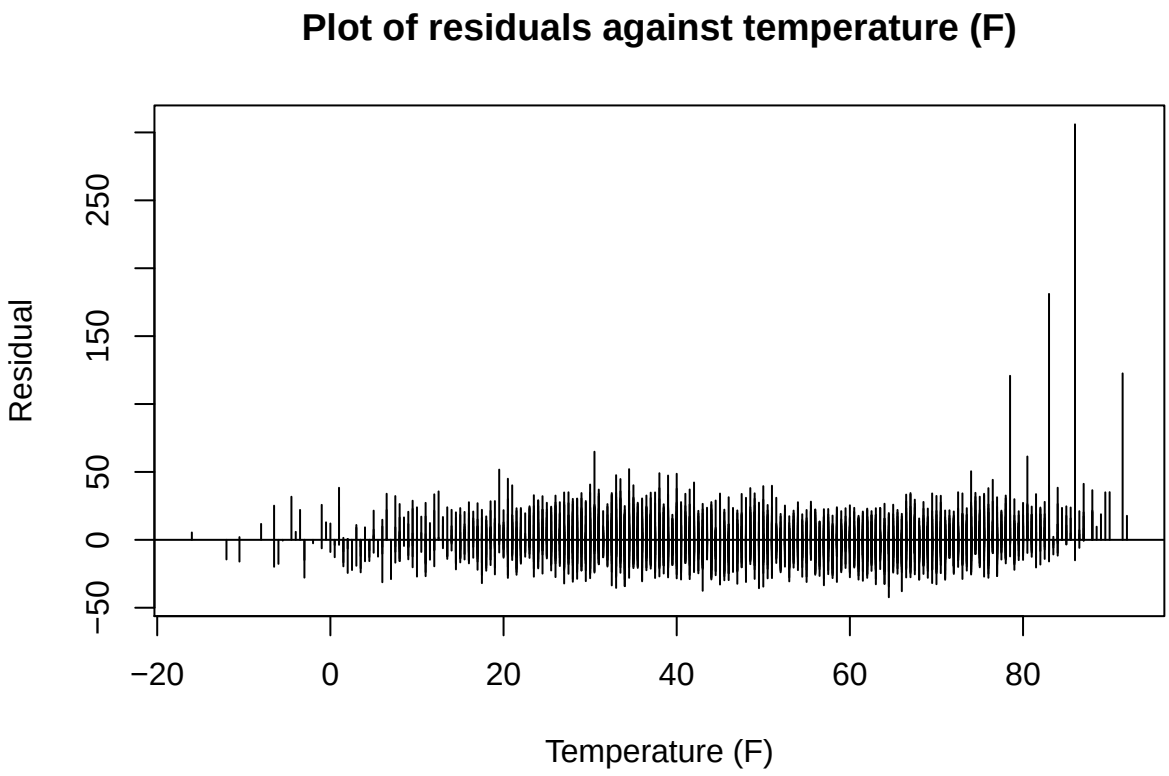
Table 3: Coefficients

	x
(Intercept)	129.957
tmpd	-0.290

Q4 c)

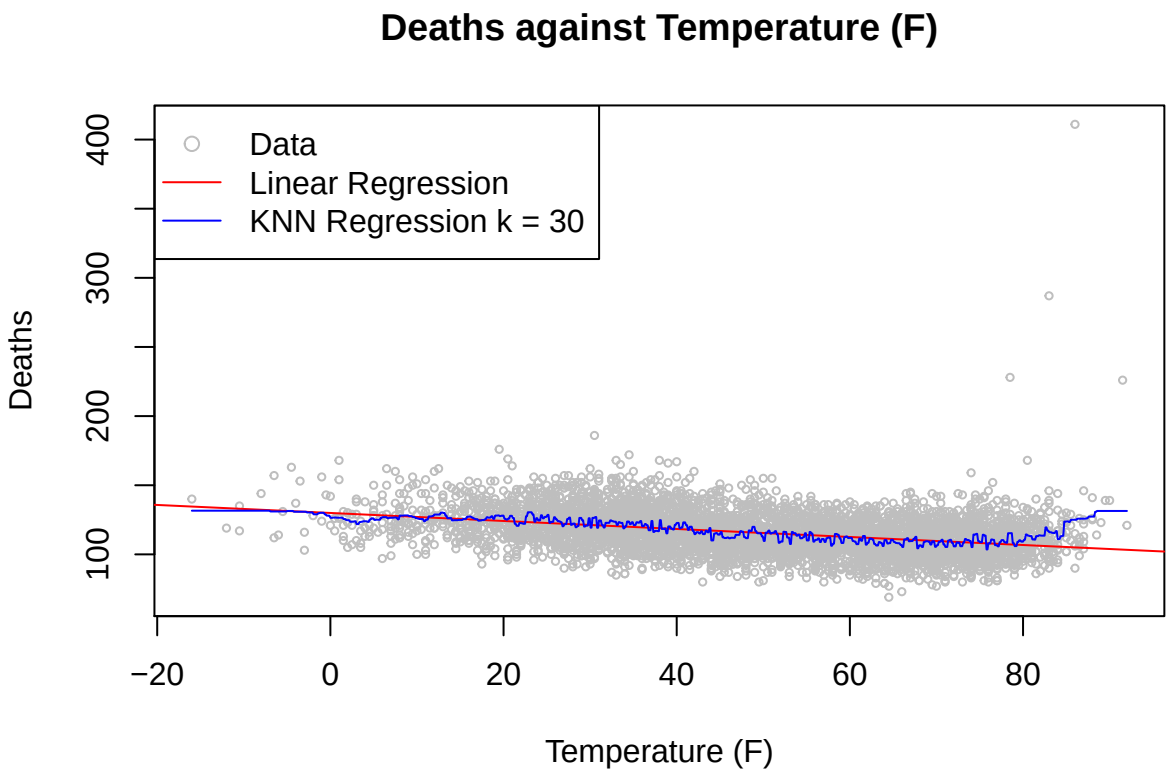
For each increase in 1 F in temperature, we predict a decrease of 0.290 deaths on average.

Q4 d)



There seems to be a rough random scatter centered around 0 for most of the residuals, with no significant signs of non-linearity. However, the outliers identified in the higher temperatures do result in large positive spikes in the residuals.

Q5 a)



Q5 b)

In comparison to the linear regression, the 30-nearest neighbour regression tracks the original data more, being roughly linear from -20F to 70F, but then increasing with data towards the higher temperatures. It probably has lower bias compared to the linear regression.

However, the 30-nearest neighbour regression is noisier than the linear regression, having higher variance overall. Even for the parts where the data seems roughly linear and has many data-points, the 30-nearest neighbour regression exhibits more rapid small changes.

Question 6

Q6 a)

```
temp.celsius <- (chicago$tmpd - 32) * (5 / 9) + 4
chicago$warmer <- temp.celsius * (9 / 5) + 32
```

Q6 b)

Table 4: Avg. change in number of deaths

x
-2.08544

Q6 c)

Table 5: Avg. change in number of deaths

x
-0.482238

Theory Problems

1) The $n \times n$ influence matrix and the degrees of freedom are:

$$\mathbf{w} = \begin{bmatrix} \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\ \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \end{bmatrix} \tag{1}$$

$$df = \text{tr}(\mathbf{w}) \tag{2}$$

$$= \sum_{i=1}^{n} \frac{1}{n} \tag{3}$$

$$= 1 \tag{4}$$

2) The $n \times n$ influence matrix and the degrees of freedom are:

$$\mathbf{w}_{ij} = \begin{cases} \frac{1}{k} & y_j \text{ is one of the } k \text{ nearest neighbors of } y_i \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

$$df = \text{tr}(\mathbf{w}) \tag{6}$$

$$= \sum_{i=1}^{n} \mathbf{w}_{ii} \tag{7}$$

$$= \frac{n}{k} \tag{8}$$

3)

$$\text{Var}\left[\hat{Y}_i\right] = \text{Var}\left[\sum_{j=1}^{n} w(x_j, x_i) Y_i\right] \tag{9}$$

$$= \sum_{j=1}^{n} \text{Var}\left[w(x_j, x_i) Y_i\right] \tag{10}$$

$$= \sum_{j=1}^{n} w^2(x_j, x_i) \text{Var}\left[Y_i\right] \tag{11}$$

$$= \sigma^2 \sum_{j=1}^{n} w^2(x_j, x_i) \tag{12}$$

$$\frac{1}{n} \sum_{i=1}^{n} \text{Var}\left[\hat{Y}_i\right] = \frac{\sigma^2}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} w^2(x_j, x_i) \tag{13}$$

$$= \frac{\sigma^2}{n} \text{tr}(\mathbf{w}\mathbf{w}^T) \tag{14}$$

In ordinary linear regression, $\mathbf{w} = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T$, which is symmetric ($\mathbf{w}^T = \mathbf{w}$) and indempotent ($\mathbf{w}^2 = \mathbf{w}$).

$$\frac{1}{n}\sum_{i=1}^{n}\text{Var}\left[\hat{Y}_i\right] = \frac{\sigma^2}{n}\text{tr}(\mathbf{w}\mathbf{w}^T) \tag{15}$$

$$= \frac{\sigma^2}{n}\text{tr}(\mathbf{w}^2) \tag{16}$$

$$= \frac{\sigma^2}{n}\text{tr}(\mathbf{w}) \tag{17}$$

$$= \frac{\sigma^2}{n}p \tag{18}$$

4)

$$\sum_{i=1}^{n}\frac{\text{Cov}\left[Y_i, \hat{Y}_i\right]}{\sigma_i^2} = \sum_{i=1}^{n}\frac{\text{Cov}\left[Y_i, \sum_{j=1}^{n} w_{ij}Y_j\right]}{\sigma_i^2} \tag{19}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\frac{\text{Cov}\left[Y_i, Y_j\right]}{\sigma_i^2} \tag{20}$$

$$= \sum_{i=1}^{n} w_{ii}\frac{\text{Var}\left[Y_i\right]}{\sigma_i^2}, \text{ as } \epsilon_i \text{ are uncorrelated} \tag{21}$$

$$= \sum_{i=1}^{n} w_{ii}\frac{\sigma_i^2}{\sigma_i^2} \tag{22}$$

$$= \sum_{i=1}^{n} w_{ii} \tag{23}$$

$$= \text{tr}(\mathbf{w}) \tag{24}$$