2.

(2) 1. Given $y = w^T \phi(x)$.

Considering gaussian distribution for each dimension of y.

let $y_i = [y_i^{(1)} \; y_i^{(2)} \dots y_i^{(k)}] \quad 1 \le i \le N$.

$\phi(x) = [a_{ij}]_{M \times N}$ where a_{ij} = i^{th} dimension value of x_j.

let $y'_p = [y_1^{(p)} \; y_2^{(p)} \dots y_N^{(p)}]^T$

$\Rightarrow y'_p | x \sim \mathcal{N}(0, \sigma^2) + \phi(x)^T w_p$

$$L = \exp\left(-\frac{1}{2\sigma^2}\left(\phi(x)^T w_p - y'_p\right)^2\right)$$

$$\log L = -\frac{1}{2\sigma^2}\left(\phi(x)^T w_p - y'_p\right)^2$$

find optimal value for w_i

$$\Rightarrow \frac{\partial \log(L)}{\partial w_i} = 0$$

$\Rightarrow$

$$\phi(x) y'_p - \phi(x)\phi(x)^T w_p = 0.$$

$$\Rightarrow \quad w_p = (\phi(x)\phi(x)^T)^{-1}\phi(x) y'_p$$

$$\Rightarrow \quad W = (\phi(x)\phi(x)^T)^{-1}\phi(x) y'$$

$\therefore$ MLE of w

$$w = (\phi(x)\,\phi(x)^T)^{-1}\,\phi(x)\,y.$$

MAP estimate:

Assuming a gaussian prior for w_i with parameter as λ.

$$p(w_i/\lambda) = \left(\frac{\lambda}{2\pi}\right)^{M/2} \exp\left(-\frac{\lambda}{2} w_i^T w\right)$$

from Baye's rule, we have

$$\underbrace{p(W/x,y,\lambda)}_{\text{posterior}} \propto \underbrace{p(x,y/w)}_{L} \cdot \underbrace{p(W/\lambda)}_{\text{prior}}$$

$$\Rightarrow p(W/x,y,\lambda) \propto p(x,y/w) \cdot \prod_{i=1}^{K} p(w_i/\lambda).$$

$$\propto \exp\left(\frac{-1}{2} \sum_{i=1}^{N} \sum_{j=1}^{k} \frac{y_i^{(j)} - w_j^T \phi(x_i)}{\sigma_j^2} \right) \exp\left(- \sum_{j=1}^{k} \frac{\lambda}{2} w_j^T w_j \right)$$

$$\Rightarrow \log\left(P(w / x, y, \lambda) \right) = \frac{-1}{2} \sum_{i=1}^{N} \sum_{j=1}^{k} \frac{y_i^{(j)} - w_j^T \phi(x_i)}{\sigma_j^2} - \sum_{j=1}^{k} \frac{\lambda}{2} w_j^T w_j + \text{Constant}$$

maximizing log-MAP to find the estimate for w_j

$$\frac{\partial \log\left(P(w / x, y, \lambda) \right)}{\partial w_j} = 0$$

$$\Rightarrow \quad -\phi(x) y_j' + \phi(x) \phi(x)^T w_j + \lambda w_j = 0.$$

$$\Rightarrow \quad w_j = \left(\phi(x) \phi(x)^T + \lambda I \right)^{-1} \phi(x) y_j'$$

MAP estimate of w

$$\Rightarrow \quad W = \left(\phi(x) \phi(x)^T + \lambda I \right)^{-1} \phi(x) y.$$

which is same as solution for ridge regression.

Q4 Considering $\phi(0) = (1,0)^T$, $\phi(1) = (0,1)^T$

$$\phi(x) = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \qquad y = w^T \phi(x)$$

$$y = \begin{bmatrix} -1 & -1 \\ -1 & -2 \\ -2 & -1 \\ 1 & 1 \\ 1 & 2 \\ 2 & 1 \end{bmatrix}$$

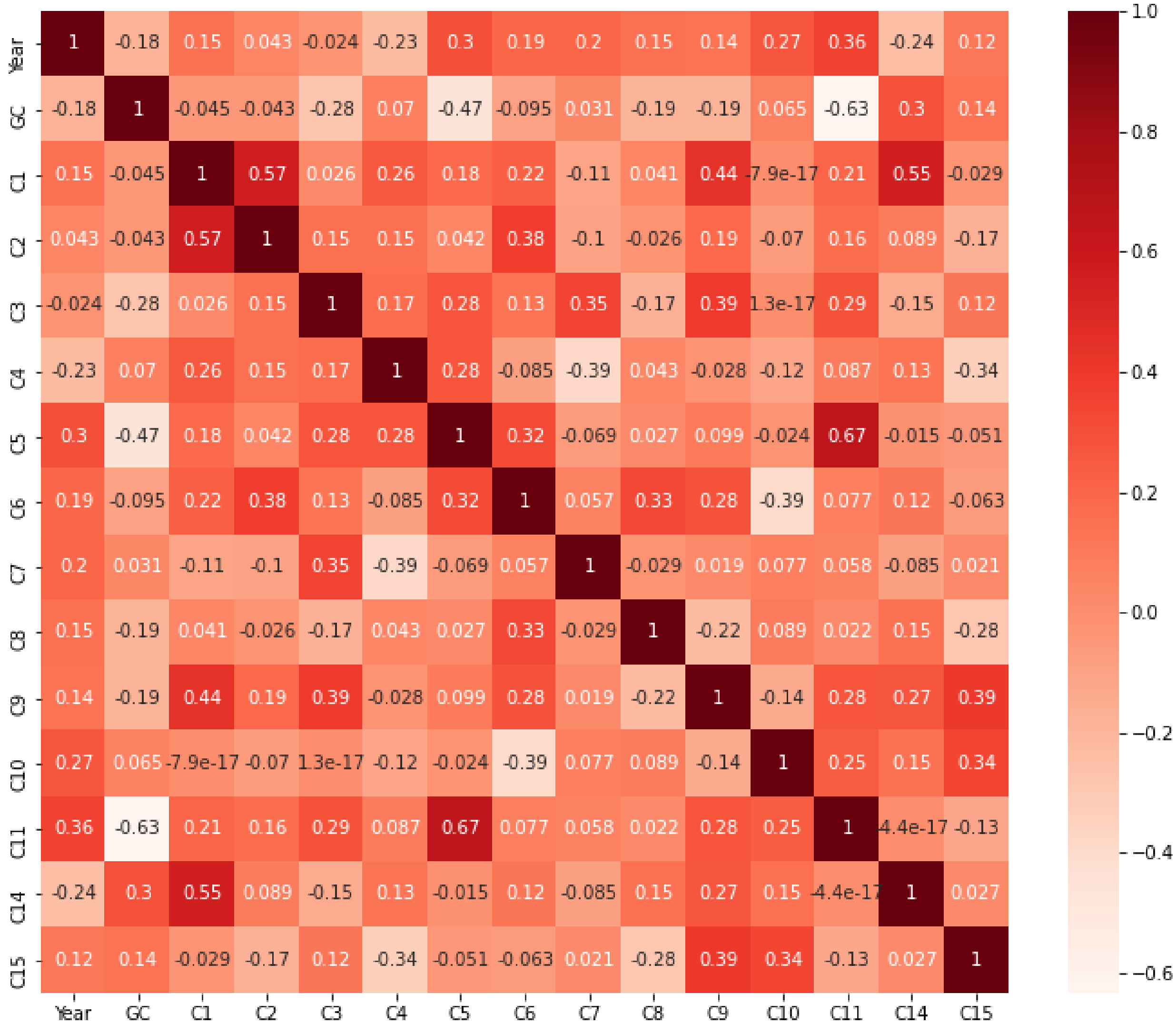
MLE of $w = (\phi(x)\phi^T(x))^{-1} \phi(x) y$

$$= \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} -1 & -1 \\ -1 & -2 \\ -2 & -1 \\ 1 & 1 \\ 1 & 2 \\ 2 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1/3 & 0 \\ 0 & 1/3 \end{bmatrix} \begin{bmatrix} -4 & -4 \\ 4 & 4 \end{bmatrix} = \begin{bmatrix} -4/3 & -4/3 \\ 4/3 & 4/3 \end{bmatrix}.$$

$$\therefore \quad w = \begin{bmatrix} -4/3 & -4/3 \\ 4/3 & 4/3 \end{bmatrix}.$$

3. Since, the number of deaths due to horse kick is not correlated with year number and the input has no other features, we will take expectation of poisson distribution as the predicted value for any input.



For proving that the feature year is not correlated we will first plot the Pearson correlation heatmap and see there isn't much correlation of corps with year. Hence, we use expectation as the predicted value.

For maximum likelihood estimation:

Given observations $y_1, y_2, \dots y_n$ for input $x_1, x_2 \dots x_n$.

$$L(\theta) = f(x_1, x_2, \dots, x_n | \theta)$$ if θ is true val of param. the probability that we observe $x_1, x_2 \dots x_n$.

for MLE, we maximize $L(\theta)$.

As our data points are iid,

maximize $L(\theta) = f(x_1|\theta) \cdot f(x_2|\theta) \dots f(x_n|\theta)$

For likelihood on poisson distribution,

$$P(Y|\lambda) = f(y_1|\lambda) \cdot f(y_2|\lambda) \dots f(y_n|\lambda)$$

$$\Rightarrow P(Y|\lambda) = \frac{e^{-n\lambda} \lambda^{\sum_{i=1}^{n} y_i}}{\prod_{i=1}^{n} y_i!} \; ; \; \left(\because f(y_i|\lambda) = \frac{e^{-\lambda} \lambda^{y_i}}{y_i!} \right)$$

& $Y = y_1, y_2, \dots y_n$.

$$\lambda_{ML} = \arg\max_{\lambda} \log (P(Y|\lambda))$$

$$\Rightarrow \frac{d}{d\lambda} \left(-n\lambda + \sum_{i=1}^{n} y_i \log\lambda + (-1) \log \left(\prod_{i=1}^{n} y_i! \right) \right) = 0$$

$$\Rightarrow \lambda_{ML} = \frac{\sum_{i=1}^{n} y_i}{n}.$$

Refer **Table 1** below for the poisson parameters (ML) and rmse values.

For maximum aposteriori estimation:

The gamma distribution is the conjugate prior for the likelihood function - poisson. Hence, we choose gamma distribution for prior distribution. It has two hyper parameters: (alpha. beta) which can be found via grid search over training set.

Gamma distribution with parameters → (α, β):

$$P(\lambda \mid \alpha, \beta) = \frac{\beta^{\alpha} \lambda^{\alpha-1} e^{-\beta\lambda}}{\Gamma(\alpha)}$$

$$P(\lambda \mid y, \alpha, \beta) \propto P(y \mid \lambda) \cdot P(\lambda \mid \alpha, \beta)$$

$$\propto \frac{e^{-(\beta+n)\lambda} \, \lambda^{\left(\sum_{i=1}^{n} y_i + \alpha - 1\right)}}{\Gamma(\alpha) \prod_{i=1}^{n} y_i!}$$

$$\lambda_{MAP} = \underset{\lambda}{\operatorname{argmax}} \log P(\lambda \mid y, \alpha, \beta)$$

$$\Rightarrow \frac{d}{d\lambda}\left(-(\beta+n)\lambda + \left(\sum_{i=1}^{n} y_i + \alpha - 1\right) \log\lambda - \log\left(\Gamma(\alpha) \prod_{i=1}^{n} y_i!\right)\right) = 0$$

$$\Rightarrow -(\beta+n) + \frac{\sum_{i=1}^{n} y_i + \alpha - 1}{\lambda} = 0 \Rightarrow \lambda = \frac{\sum_{i=1}^{n} y_i + \alpha - 1}{n + \beta}$$

After doing a grid search over possible values of (alpha, beta) from {1..10} and taking Σ(rmse) as the cost metric over all corps on the training set, we find that alpha = 2 and beta = 1 is the best fit for it.

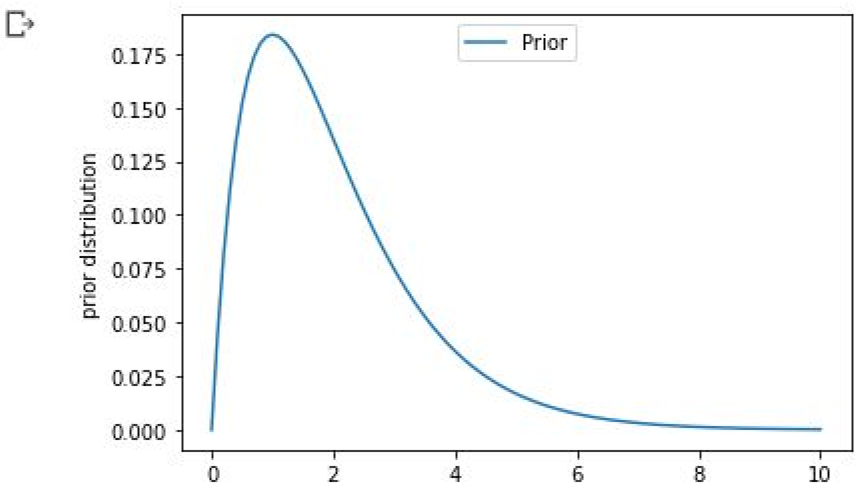
Finally **Table 1** below shows the MAP parameters and rmse values.

Corp	**ML**		**MAP**	
	λ	RMSE	λ	RMSE
G	1	0.755929	1	0.755929
I	0.692308	1.11244	0.714286	1.10657
II	0.615385	0.729756	0.642857	0.731925
III	0.615385	0.729756	0.642857	0.731925
IV	0.461538	0.484764	0.5	0.5
V	0.384615	0.587989	0.428571	0.553283
VI	0.846154	0.989804	0.857143	0.989743
VII	0.538462	0.898011	0.571429	0.892143
VIII	0.307692	0.509421	0.357143	0.5
IX	0.692308	0.738393	0.714286	0.742307
X	0.538462	1.15969	0.571429	1.14286
XI	1	1.13389	1	1.13389
XIV	1.46154	1.02381	1.42857	1
XV	0.307692	0.941214	0.357143	0.928571

Table 1

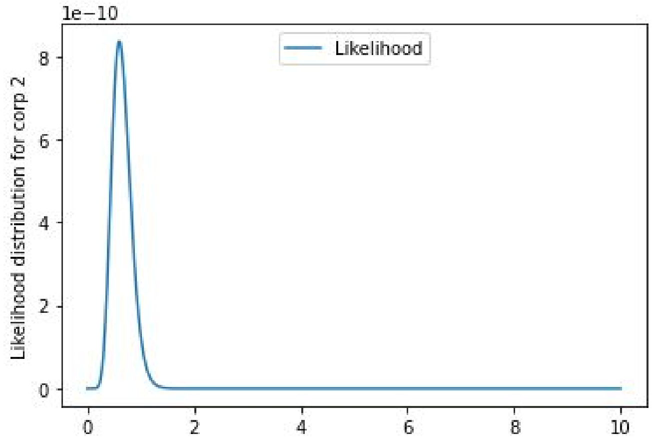
Graphs

Prior

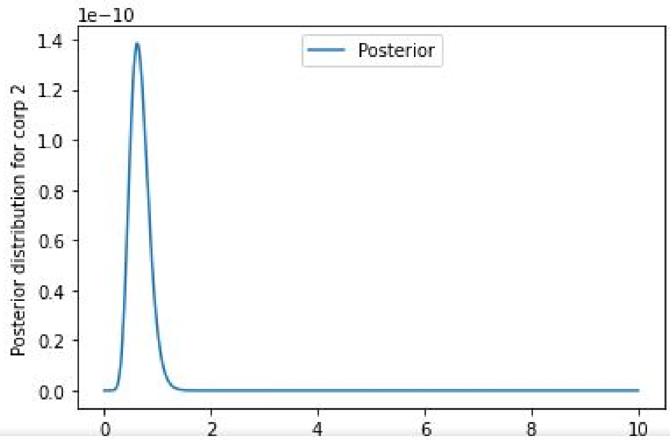



```
(Prior)Maxima at:  1.0
```

Corp 2

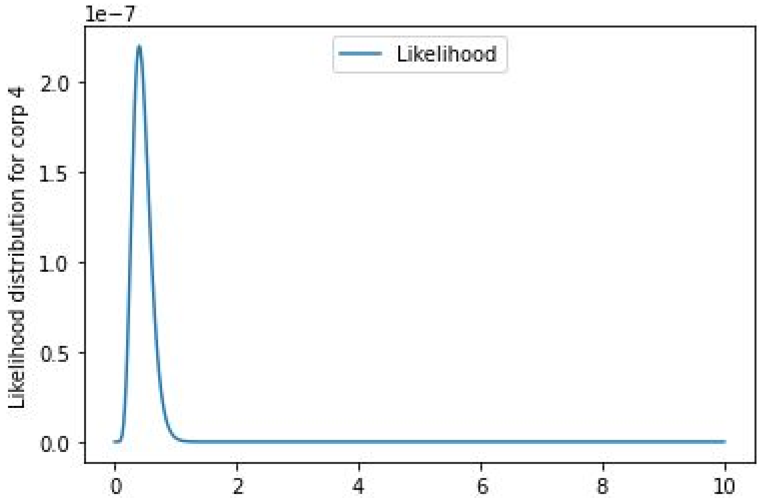



```
(Likelihood)Maxima at:  0.6
```

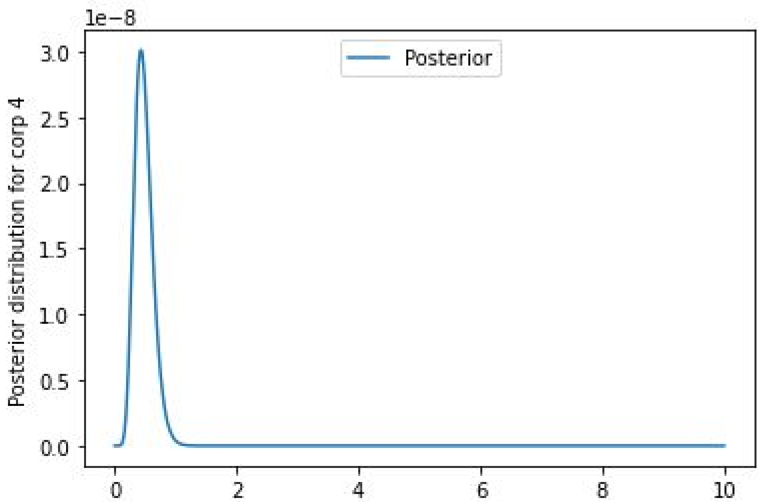



```
(Posterior)Maxima at:  0.62
```

Corp 4

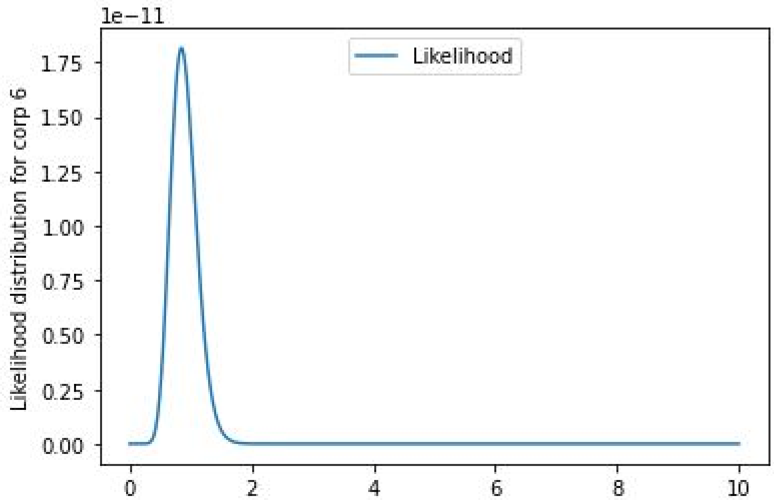



```
(Likelihood)Maxima at:  0.4
```

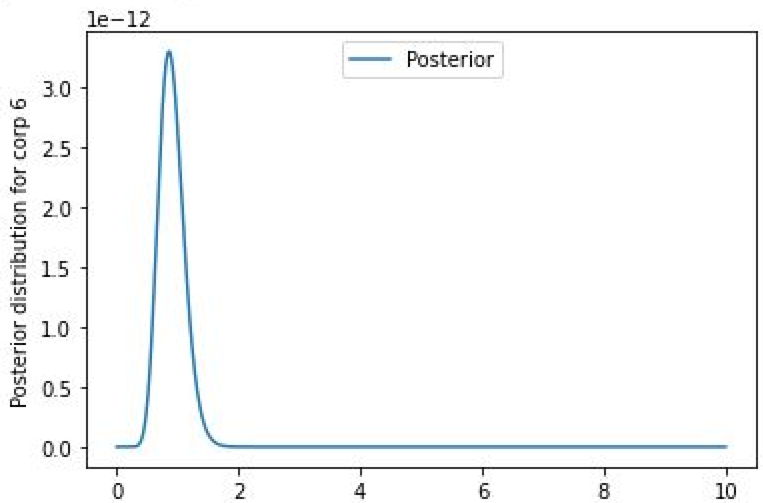



```
(Posterior)Maxima at:  0.43
```

Corp 6



(Likelihood)Maxima at: 0.85



(Posterior)Maxima at: 0.86

The above maximas are obtained after considering the whole dataset for Likelihood and Prior graphs. The graph has been plotted against lambda values with precision 0.1, Maximas for which have been written below the corresponding ones.