1. a) The function to be minimized is

$$f(m) = \frac{1}{n}\sum_{i=1}^{n} |Y_i - m|^2 = \frac{1}{n}\sum_{i=1}^{n} (Y_i - m)^2$$

$f(m)$ is a polynomial in m so it is differentiable on $\mathbb{R}$ so the minimum of f occurs at a critical point. The derivative of f is

$$\begin{aligned}
\frac{df}{dm} &= \frac{d}{dm}\frac{1}{n}\sum_{i=1}^{n}(Y_i - m)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{d}{dm}(Y_i - m)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} 2(Y_i - m)(-1) \\
&= -\frac{2}{n}\left[\left(\sum_{i=1}^{n} Y_i\right) - \left(\sum_{i=1}^{n} m\right)\right]
\end{aligned}$$

Setting $\dfrac{df}{dm} = 0$ gives

$$\begin{aligned}
-\frac{2}{n}\left[\left(\sum_{i=1}^{n} Y_i\right) - \left(\sum_{i=1}^{n} m\right)\right] &= 0 \\
\Rightarrow \left(\sum_{i=1}^{n} Y_i\right) - \left(\sum_{i=1}^{n} m\right) &= 0 \\
\Rightarrow \sum_{i=1}^{n} Y_i &= nm \\
\Rightarrow m &= \frac{1}{n}\sum_{i=1}^{n} Y_i \\
\Rightarrow m &= h_{\text{avg}}
\end{aligned}$$

The second derivative of f is

$$\begin{aligned}
\frac{d^2 f}{dm^2} &= \frac{d}{dm}\left[-\frac{2}{n}\left[\left(\sum_{i=1}^{n} Y_i\right) - \left(\sum_{i=1}^{n} m\right)\right]\right] \\
&= -\frac{2}{n}\left[-\left(\sum_{i=1}^{n} 1\right)\right] \\
&= 2 \\
&> 0
\end{aligned}$$

so by the second derivative test, h_{avg} is a local minimum of f. Since it is the only critical point of f, $m = h_{\text{avg}}$ is the minimum of f.

b) The bias of $h_{\text{avg}}(\mathcal{D})$ is

$$\begin{aligned}
|\mathbb{E}_{\mathcal{D}}[h_{\text{avg}}(\mathcal{D})] - \mu|^2 &= \left| \mathbb{E}_{\mathcal{D}} \left[\frac{1}{n} \sum_{i=1}^{n} Y_i \right] - \mu \right|^2 \\
&= \left| \frac{1}{n} \mathbb{E}_{\mathcal{D}} \left[\sum_{i=1}^{n} Y_i \right] - \mu \right|^2 \\
&= \left| \left(\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\mathcal{D}}[Y_i] \right) - \mu \right|^2 \\
&= \left| \left(\frac{1}{n} \sum_{i=1}^{n} \mu \right) - \mu \right|^2 \\
&= |\mu - \mu|^2 \\
&= 0
\end{aligned}$$

The variance of $h_{\text{avg}}(\mathcal{D})$ is

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}[|h_{\text{avg}}(\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[h_{\text{avg}}(\mathcal{D})]|^2] &= \mathbb{E}_{\mathcal{D}} \left[\left| \left(\frac{1}{n} \sum_{i=1}^{n} Y_i \right) - \mathbb{E}_{\mathcal{D}} \left[\frac{1}{n} \sum_{i=1}^{n} Y_i \right] \right|^2 \right] \\
&= \mathbb{E}_{\mathcal{D}} \left[\left| \frac{1}{n} \left[\left(\sum_{i=1}^{n} Y_i \right) - \mathbb{E}_{\mathcal{D}} \left[\sum_{i=1}^{n} Y_i \right] \right] \right|^2 \right] \\
&= \frac{1}{n^2} \mathbb{E}_{\mathcal{D}} \left[\left| \left(\sum_{i=1}^{n} Y_i \right) - n\mu \right|^2 \right] \\
&= \frac{1}{n^2} \mathbb{E}_{\mathcal{D}} \left[\left| \sum_{i=1}^{n} (Y_i - \mu) \right|^2 \right] \\
&= \frac{1}{n^2} \sum_{i=1}^{n} \mathbb{E}_{\mathcal{D}} \left[|Y_i - \mu|^2 \right] \\
&= \frac{1}{n^2} \sum_{i=1}^{n} \mathbb{E}_{\mathcal{D}} \left[(Y_i - \mu)^2 \right] \\
&= \frac{1}{n^2} \sum_{i=1}^{n} \sigma^2 \\
&= \frac{1}{n^2} (n\sigma^2) \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

c) The function to be minimized is

$$f(m) = \frac{1}{n} \left(\sum_{i=1}^{n} |Y_i - m|^2 \right) + \lambda |m|^2 = \frac{1}{n} \left(\sum_{i=1}^{n} (Y_i - m)^2 \right) + \lambda m^2$$

$f(m)$ is a polynomial in m so it is differentiable on $\mathbb{R}$ so the minimum of f occurs at a critical point. The derivative of f is

$$\begin{aligned}
\frac{df}{dm} &= \frac{d}{dm}\left[\frac{1}{n}\left(\sum_{i=1}^{n}(Y_i - m)^2\right) + \lambda m^2\right] \\
&= \frac{1}{n}\left(\sum_{i=1}^{n}\frac{d}{dm}(Y_i - m)^2\right) + \frac{d}{dm}(\lambda m^2) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n}2(Y_i - m)(-1)\right) + 2\lambda m \\
&= -\frac{2}{n}\left[\left(\sum_{i=1}^{n}Y_i\right) - \left(\sum_{i=1}^{n}m\right)\right] + 2\lambda m
\end{aligned}$$

Setting $\frac{df}{dm} = 0$ gives

$$\begin{aligned}
-\frac{2}{n}\left[\left(\sum_{i=1}^{n}Y_i\right) - \left(\sum_{i=1}^{n}m\right)\right] + 2\lambda m = 0 \\
\Rightarrow \frac{2}{n}\left(\sum_{i=1}^{n}m\right) + 2\lambda m = \frac{2}{n}\sum_{i=1}^{n}Y_i \\
\Rightarrow \frac{2}{n}(nm) + 2\lambda m = \frac{2}{n}\sum_{i=1}^{n}Y_i \\
\Rightarrow m(1+\lambda) = \frac{1}{n}\sum_{i=1}^{n}Y_i \\
\Rightarrow m = \frac{h_{\text{avg}}}{\lambda + 1}
\end{aligned}$$

The second derivative of f is

$$\begin{aligned}
\frac{d^2f}{dm^2} &= \frac{d}{dm}\left[-\frac{2}{n}\left[\left(\sum_{i=1}^{n}Y_i\right) - \left(\sum_{i=1}^{n}m\right)\right] + 2\lambda m\right] \\
&= -\frac{2}{n}\left[-\left(\sum_{i=1}^{n}1\right)\right] + 2\lambda \\
&= 2 + 2\lambda \\
&> 0
\end{aligned}$$

since $\lambda \geq 0$. By the second derivative test, $\frac{h_{\text{avg}}}{\lambda + 1}$ is a local minimum of f. Since it is the only critical point of f, $h_\lambda(\mathcal{D}) = \frac{h_{\text{avg}}}{\lambda + 1}$ is the minimum of f.

d) The bias of $h_\lambda(\mathcal{D})$ is

$$\begin{aligned}
|\mathbb{E}_{\mathcal{D}}[h_\lambda(\mathcal{D})] - \mu|^2 &= \left| \mathbb{E}_{\mathcal{D}} \left[\frac{h_{\text{avg}}}{\lambda+1} \right] - \mu \right|^2 \\
&= \left| \frac{1}{n(\lambda+1)} \mathbb{E}_{\mathcal{D}} \left[\sum_{i=1}^{n} Y_i \right] - \mu \right|^2 \\
&= \left| \left(\frac{1}{n(\lambda+1)} \sum_{i=1}^{n} \mathbb{E}_{\mathcal{D}}[Y_i] \right) - \mu \right|^2 \\
&= \left| \frac{1}{\lambda+1} \left(\frac{1}{n} \sum_{i=1}^{n} \mu \right) - \mu \right|^2 \\
&= \left| \frac{\mu}{\lambda+1} - \mu \right|^2 \\
&= \left| \frac{\mu}{\lambda+1} - \frac{\mu(\lambda+1)}{\lambda+1} \right|^2 \\
&= \left| \frac{-\mu\lambda}{\lambda+1} \right|^2 \\
&= \frac{\mu^2\lambda^2}{(\lambda+1)^2}
\end{aligned}$$

The variance of $h_\lambda(\mathcal{D})$ is

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}[|h_\lambda(\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[h_\lambda(\mathcal{D})]|^2] &= \mathbb{E}_{\mathcal{D}}\left[\left|\left(\frac{h_{\text{avg}}}{\lambda+1}\right) - \mathbb{E}_{\mathcal{D}}\left[\frac{h_{\text{avg}}}{\lambda+1}\right]\right|^2\right] \\
&= \mathbb{E}_{\mathcal{D}}\left[\left|\frac{1}{n(\lambda+1)}\left[\left(\sum_{i=1}^n Y_i\right) - \mathbb{E}_{\mathcal{D}}\left[\sum_{i=1}^n Y_i\right]\right]\right|^2\right] \\
&= \frac{1}{n^2(\lambda+1)^2}\mathbb{E}_{\mathcal{D}}\left[\left|\left(\sum_{i=1}^n Y_i\right) - n\mu\right|^2\right] \\
&= \frac{1}{n^2(\lambda+1)^2}\mathbb{E}_{\mathcal{D}}\left[\left|\sum_{i=1}^n (Y_i - \mu)\right|^2\right] \\
&= \frac{1}{n^2(\lambda+1)^2}\sum_{i=1}^n \mathbb{E}_{\mathcal{D}}\left[|Y_i - \mu|^2\right] \\
&= \frac{1}{n^2(\lambda+1)^2}\sum_{i=1}^n \mathbb{E}_{\mathcal{D}}\left[(Y_i - \mu)^2\right] \\
&= \frac{1}{n^2(\lambda+1)^2}\sum_{i=1}^n \sigma^2 \\
&= \frac{1}{n^2(\lambda+1)^2}(n\sigma^2) \\
&= \frac{\sigma^2}{n(\lambda+1)^2}
\end{aligned}$$

e) Using $\mu = 1, \sigma^2 = 9$, and $n = 10$, the bias simplifies to

$$\begin{aligned}
|\mathbb{E}_{\mathcal{D}}[h_\lambda(\mathcal{D})] - \mu|^2 &= \frac{\mu^2\lambda^2}{(\lambda+1)^2} \\
&= \frac{(1^2)\lambda^2}{(\lambda+1)^2} \\
&= \frac{\lambda^2}{(\lambda+1)^2}
\end{aligned}$$

The variance simplifies to

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}[|h_\lambda(\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[h_\lambda(\mathcal{D})]|^2] &= \frac{\sigma^2}{n(\lambda+1)^2} \\
&= \frac{9}{10(\lambda+1)^2}
\end{aligned}$$

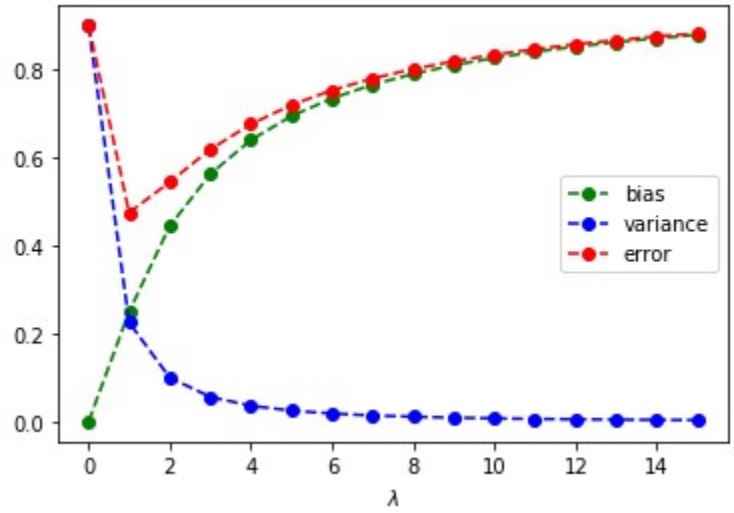
By the bias-variance decomposition and d),

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}[|h_\lambda(\mathcal{D}) - \mu|^2] &= |\mathbb{E}_{\mathcal{D}}[h_\lambda(\mathcal{D})] - \mu|^2 + \mathbb{E}_{\mathcal{D}}[|h_\lambda(\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[h_\lambda(\mathcal{D})]|^2] \\ &= \frac{\mu^2\lambda^2}{(\lambda+1)^2} + \frac{\sigma^2}{n(\lambda+1)^2} \\ &= \frac{n\mu^2\lambda^2}{n(\lambda+1)^2} + \frac{\sigma^2}{n(\lambda+1)^2} \\ &= \frac{n\mu^2\lambda^2 + \sigma^2}{n(\lambda+1)^2}\end{aligned}$$

so

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}[|h_\lambda(\mathcal{D}) - \mu|^2] &= \frac{10(1^2)\lambda^2 + 9}{10(\lambda+1)^2} \\ &= \frac{10\lambda^2 + 9}{10(\lambda+1)^2}\end{aligned}$$

Plotting the three equations gives the following graph.



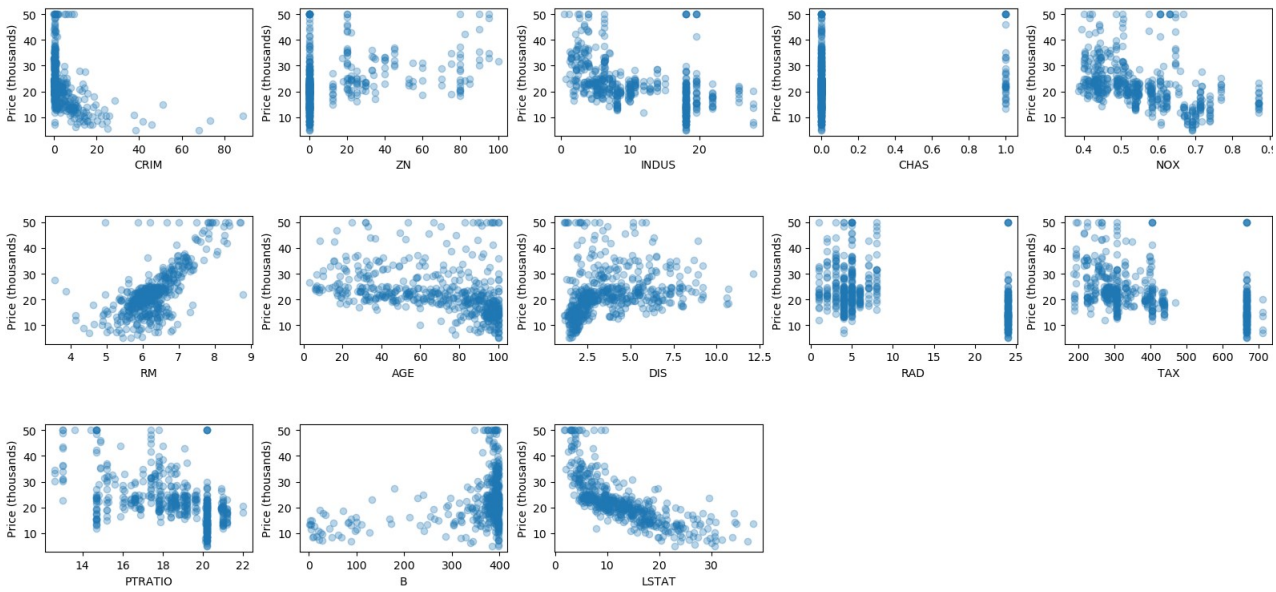
f) As λ increases, the variance goes down while the bias goes up. When λ goes from 0 to 1, the drop in variance is much greater than the increase in bias, which causes the error, which is the sum of the bias and variance, to drop substantially. Afterwards, increases in bias and decreases in variance when λ increases both get smaller and become more and more similar in magnitude. The magnitude of the bias increase is always slightly more than the magnitude of the variance decrease, which causes the error to move slightly upwards as λ increases to beyond 1. Using the bias and variance equations from e) and setting $\lambda \to \infty$, it is seen that the variance approaches 0 while the bias approach 0.9 so the error approaches $0 + 0.9 = 0.9$ as λ increases.

2. All parts have related code in the code file.

b)

```
Number of data points: 506
Number of features: 13
Correlation coefficient between CRIM and price: -0.3883046085868114
Correlation coefficient between ZN and price: 0.3604453424505444
Correlation coefficient between INDUS and price: -0.48372516002837346
Correlation coefficient between CHAS and price: 0.17526017719029738
Correlation coefficient between NOX and price: -0.42732077237328137
Correlation coefficient between RM and price: 0.6953599470715387
Correlation coefficient between AGE and price: -0.3769545650045959
Correlation coefficient between DIS and price: 0.24992873408590385
Correlation coefficient between RAD and price: -0.38162623063977735
Correlation coefficient between TAX and price: -0.4685359335677663
Correlation coefficient between PTRATIO and price: -0.507786685537561
Correlation coefficient between B and price: 0.33346081965706603
Correlation coefficient between LSTAT and price: -0.7376627261740145
Mean house price: 22532.81 dollars
Median house price: 21200.0 dollars
Third quartile house price: 25000.0 dollars
First quartile house price: 17025.0 dollars
House price IQR 7975.0 dollars
House price variance: 84419556.16 square dollars
House price standard deviation: 9188.01 dollars
House price mode: 50000.0 dollars
House price maximum: 50000.0 dollars
House price minimum: 5000.0 dollars
House price range: 45000.0 dollars
```

c)



e)

Fitted weights:

Bias	CRIM	ZN	INDUS	CHAS	NOX	RM
18.8603	-0.00386393	0.0309677	0.170872	2.64437	-20.0296	5.68337

AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
-0.014947	-1.42542	0.27274	-0.0129428	-0.779786	0.0166983	-0.436918

The weight in the third weight column (INDUS) has a positive sign, which indicates there is a positive correlation associated with the attribute (proportion of non-retail business acres per town) and house value.

The sign does not match what I expected. I expected that non-retail businesses encompasses business that perform resource extraction and manufacturing, which indicates land is for farming and is far from urban centers. Such land would be far away from businesses, schools, medical care, and other facilities typical of urban areas so I would expect that as the proportion of such land increases, the house price decreases which would mean a negative weight sign.

f) The first error metric is the mean absolute error, which given $(a_1, p_1), \ldots (a_n, p_n)$ of pairs of actual and predicted values, the error is

$$E_{\text{MAE}} = \frac{1}{n} \sum_{i=1}^{n} |a_n - p_n|$$

The second error metric is the maximum error, which given $(a_1, p_1), \ldots (a_n, p_n)$ of pairs of actual and predicted values, the error is

$$E_{\text{ME}} = \max\{|a_i - p_i|\}_{i=1}^{n}$$

```
Mean squared error:  28.485690516004446
```

g)

```
Mean absolute error:  3.7601448170106493

Max error:  24.430139416598053
```

h) The features that best predict price are those that show the strongest correlation with price. These features have the greatest correlation coefficient magnitude with price. From the correlation coefficient calculations in image in b), the two most significant features are thus LSTAT and RM, followed by PTRATIO, INDUS, and TAX.

3a) Let

$$f(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N} a^{(i)}(y^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)})^2$$

and

$$g(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N}(y^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)})^2$$

$g(x)$ is half the squared norm of the vector of length N with component i equal to $y^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)}$, which is $\mathbf{y} - \mathbf{X}\mathbf{w}$ so $g(\mathbf{w}) = \frac{1}{2}||\mathbf{y} - \mathbf{X}\mathbf{w}||^2$.

Let $\sqrt{\mathbf{A}}$ be the $N \times N$ diagonal matrix with $\sqrt{\mathbf{A}}_{ii} = \sqrt{a^{(i)}}$ for all i which is defined since $a^{(i)} > 0$ for all i. $f(\mathbf{w})$ multiplies $a^{(i)}$ to each component i which can be introduced by multiplying the i the component of $\mathbf{y} - \mathbf{X}\mathbf{w}$ by $\sqrt{a^{(i)}}$ in the expression of $g(\mathbf{w})$. This can be written as $\sqrt{\mathbf{A}}(\mathbf{y} - \mathbf{X}\mathbf{w}) = \sqrt{\mathbf{A}}\mathbf{y} - \sqrt{\mathbf{A}}\mathbf{X}\mathbf{w}$ so $f(\mathbf{w}) = \frac{1}{2}||\sqrt{\mathbf{A}}\mathbf{y} - \sqrt{\mathbf{A}}\mathbf{X}\mathbf{w}||^2$.

If $\mathbf{a} = [a_1, \ldots, a_N]$ and $\mathbf{b} = [b_1, \ldots, b_N]$ are two vectors of length N, then

$$\begin{aligned}
||\mathbf{a} - \mathbf{b}||^2 &= \sum_{i=1}^{N}(a_i - b_i)^2 \\
&= \sum_{i=1}^{N}(a_i^2 - 2a_ib_i + b_i^2) \\
&= \sum_{i=1}^{N} a_i^2 - 2\sum_{i=1}^{N} a_ib_i + \sum_{i=1}^{N} b_i^2 \\
&= ||\mathbf{a}||^2 - 2\mathbf{a}\cdot\mathbf{b} + ||\mathbf{b}||^2
\end{aligned}$$

Using this identity with $\mathbf{a} = \sqrt{\mathbf{A}}\mathbf{y}$ and $\mathbf{b} = \sqrt{\mathbf{A}}\mathbf{X}\mathbf{w}$ gives

$$\begin{aligned}
f(\mathbf{w}) &= \frac{1}{2}||\sqrt{\mathbf{A}}\mathbf{y} - \sqrt{\mathbf{A}}\mathbf{X}\mathbf{w}||^2 \\
&= \frac{1}{2}(||\sqrt{\mathbf{A}}\mathbf{y}||^2 - (\sqrt{\mathbf{A}}\mathbf{y})\cdot(\sqrt{\mathbf{A}}\mathbf{X}\mathbf{w}) + ||\sqrt{\mathbf{A}}\mathbf{X}\mathbf{w}||^2) \\
&= \frac{1}{2}(\sqrt{\mathbf{A}}\mathbf{y})^T(\sqrt{\mathbf{A}}\mathbf{y}) - (\sqrt{\mathbf{A}}\mathbf{y})^T(\sqrt{\mathbf{A}}\mathbf{X}\mathbf{w}) + \frac{1}{2}(\sqrt{\mathbf{A}}\mathbf{X}\mathbf{w})^T(\sqrt{\mathbf{A}}\mathbf{X}\mathbf{w}) \\
&= \frac{1}{2}\mathbf{y}^T\sqrt{\mathbf{A}}^T\sqrt{\mathbf{A}}\mathbf{y} - \mathbf{y}^T\sqrt{\mathbf{A}}^T\sqrt{\mathbf{A}}\mathbf{X}\mathbf{w} + \frac{1}{2}\mathbf{w}\mathbf{X}^T\sqrt{\mathbf{A}}^T\sqrt{\mathbf{A}}\mathbf{X}\mathbf{w}
\end{aligned}$$

Since $\sqrt{\mathbf{A}}$ is a diagonal matrix, we have $\sqrt{\mathbf{A}}\sqrt{\mathbf{A}} = \mathbf{A}$ and $\sqrt{\mathbf{A}} = \sqrt{\mathbf{A}}^T$ so $\sqrt{\mathbf{A}}^T\sqrt{\mathbf{A}} = \mathbf{A}$. Substituting this into the expression above gives

$$f(\mathbf{w}) = \frac{1}{2}\mathbf{y}^T\mathbf{A}\mathbf{y} - \mathbf{y}^T\mathbf{A}\mathbf{X}\mathbf{w} + \frac{1}{2}\mathbf{w}\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{w}$$

$f(\mathbf{w})$ is minimized at a point where the gradient is equal to $\mathbf{0}$. Using the property $\nabla(\mathbf{w}^T\mathbf{v}) = \mathbf{v}$ on the second term and $\nabla(\mathbf{wMw}) = 2\mathbf{Mw}$ on the third term when taking the gradient of $f(\mathbf{w})$ gives

$$\begin{aligned}\nabla f(\mathbf{w}) &= \nabla\left(\frac{1}{2}\mathbf{y}^T\mathbf{A}\mathbf{y}\right) - \nabla(\mathbf{w}^T(\mathbf{X}^T\mathbf{A}^T\mathbf{y})) + \nabla\left(\frac{1}{2}\mathbf{w}(\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{w})\right) \\ &= -\mathbf{X}^T\mathbf{A}^T\mathbf{y} + \mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{w}\end{aligned}$$

Setting $\nabla f(\mathbf{w}^*) = \mathbf{0}$ gives

$$\begin{aligned}-\mathbf{X}^T\mathbf{A}^T\mathbf{y} + \mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{w}^* &= \mathbf{0} \\ \Rightarrow (\mathbf{X}^T\mathbf{A}\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{A}\mathbf{X})\mathbf{w}^* &= (\mathbf{X}^T\mathbf{A}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{A}^T\mathbf{y} \\ \Rightarrow \mathbf{w}^* &= (\mathbf{X}^T\mathbf{A}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{A}^T\mathbf{y}\end{aligned}$$

Since $\mathbf{A}$ is diagonal, $\mathbf{A} = \mathbf{A}^T$ so $\mathbf{w}^*$ can be rewritten as

$$\mathbf{w}^* = (\mathbf{X}^T\mathbf{A}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{A}\mathbf{y}$$

The ith row of $\nabla f(\mathbf{w})$ is the dot product of the ith row of $\mathbf{X}^T\mathbf{A}\mathbf{X}$ and $\mathbf{w}$. The partial derivative with respect to w_j of this row is j element of the $i$row of $\mathbf{X}^T\mathbf{A}\mathbf{X}$. Thus, the Hessian matrix H_f of $f(\mathbf{w})$ is $\mathbf{X}^T\mathbf{A}\mathbf{X}$. Given any $\mathbf{v} \in \mathbb{R}^N$ with $\mathbf{v} \neq \mathbf{0}$, we have

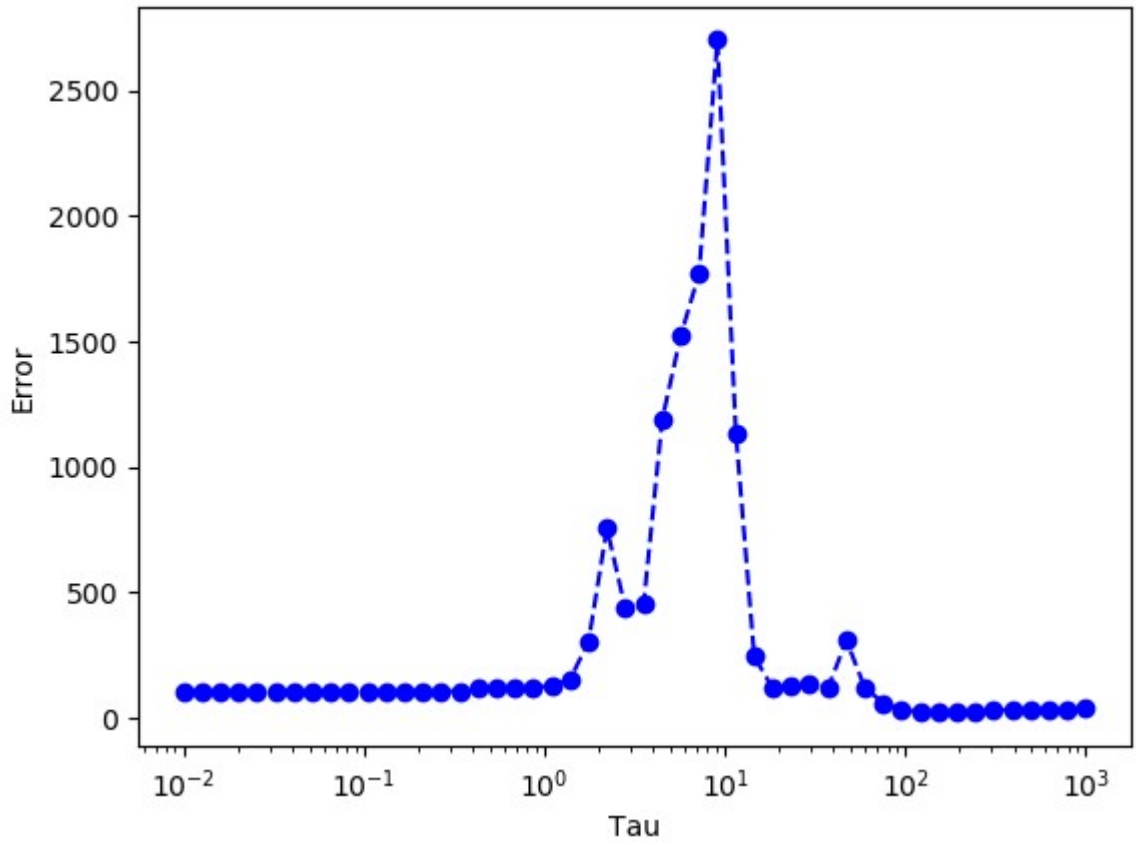
$$\begin{aligned}\mathbf{v}^T H_f \mathbf{v} &= \mathbf{v}^T(\mathbf{X}^T\mathbf{A}\mathbf{X})\mathbf{v} \\ &= (\mathbf{X}\mathbf{v})^T\mathbf{A}(\mathbf{X}\mathbf{v})\end{aligned}$$

Let $\mathbf{X}\mathbf{v} = \mathbf{z}$ and let the ith column of $\mathbf{z}$ be z_i. Let the ith diagonal element of $\mathbf{A}$ be a_i. Then

$$\begin{aligned}\mathbf{v}^T H_f \mathbf{v} &= \mathbf{z}^T\mathbf{A}\mathbf{z} \\ &= \sum_{i=1}^{N} a_i(z_i^2) \\ &> 0\end{aligned}$$

since a_i is positive for all i. Thus, H_f is positive definite so $\mathbf{w}^i$ is a local minimum of f. Since this is the only critical point of $f(\mathbf{w})$, $\mathbf{w}^* = (\mathbf{X}^T\mathbf{A}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{A}\mathbf{y}$ minimizes $f(\mathbf{w})$.

c)




```
min loss = 20.943413381683342
```

d) When $\tau \to \infty$, we have $2\tau^2 \to \infty$ so the expression $\dfrac{-||\mathbf{x} - \mathbf{x}^{(i)}||^2}{2\tau^2} \to 0$ for all i. Thus, $\exp\left(\dfrac{-||\mathbf{x} - \mathbf{x}^{(i)}||^2}{2\tau^2}\right) \to e^0 = 1$ so $a^{(i)} = \dfrac{\exp\left(\dfrac{-||\mathbf{x} - \mathbf{x}^{(i)}||^2}{2\tau^2}\right)}{\sum_j \exp\left(\dfrac{-||\mathbf{x} - \mathbf{x}^{(j)}||^2}{2\tau^2}\right)} \to \dfrac{1}{\sum_j 1} = \dfrac{1}{N}$ so the weights approach equality as $\tau \to \infty$. This shows the algorithm approaches the non-weight least squares solution when $\tau \to \infty$.

For $\tau \to 0$, let $f(x) = \exp\left(-\dfrac{x}{2\tau^2}\right)$. Given $x_1 > x_2$, we have $x_2 - x_1 < 0$ so

$$\begin{aligned}
\lim_{\tau \to 0^+} \frac{f(x_1)}{f(x_2)} &= \lim_{\tau \to 0} \frac{\exp\left(-\dfrac{x_1}{2\tau^2}\right)}{\exp\left(-\dfrac{x_2}{2\tau^2}\right)} \\
&= \lim_{\tau \to 0^+} \exp\left(\frac{x_2 - x_1}{2\tau^2}\right) \\
&= \lim_{x \to -\infty} e^x \\
&= 0
\end{aligned}$$

Thus, the term that dominates in the denominator of $a^{(i)} = \dfrac{\exp\left(\dfrac{-||\mathbf{x}-\mathbf{x}^{(i)}||^2}{2\tau^2}\right)}{\sum_j \exp\left(\dfrac{-||\mathbf{x}-\mathbf{x}^{(j)}||^2}{2\tau^2}\right)}$ when $\tau \to 0$, is the one with the smallest value of $-||\mathbf{x}-\mathbf{x}^{(j)}||^2$, which is the one with the largest value of $||\mathbf{x}-\mathbf{x}^{(j)}||^2$. For this particular k, we have $\lim_{\tau\to 0} a^{(k)} = \dfrac{1}{1+\sum_{j\neq k} 0} = 1$. For $i \neq k$, we have $\lim_{\tau\to 0} a^{(i)} = \dfrac{0}{1+\sum_{j\neq k} 0} = 0$. Thus, only the value of the term associated with the weight $a^{(k)}$ matters so the problems reduces to minimizing $\dfrac{1}{2}(y^{(k)}-\mathbf{w}^T\mathbf{x}^{(k)})^2$, which occurs when $y^{(k)}-\mathbf{w}^T\mathbf{x}^{(k)} = 0$, which can be achieved by an infinite number of $\mathbf{w}$. Since $||\mathbf{x}-\mathbf{x}^{(j)}||^2$ is maximized when $\mathbf{x}^{(j)}$ is far from $\mathbf{x}$, the algorithm reduces to the non-weight least squares problem with only the input that is furthest from $\mathbf{x}$ and the associated target.

e) Advantage 1: Locally weighted linear regression is a local algorithm so it tunes its weights by focusing on the input data point rather than fixing weight across multiple data points like ordinary linear regression. This makes locally weight linear regression accurate on the input data point as it depends less on previous data points compared to non-weight linear regression.

Advantage 2: As the calculations in part e) show, training examples that are far away from the input data point have much greater impact on the algorithm output than training examples that are closer to the input data point when $\tau \to 0$ and the training examples all have the same weight (ordinary linear regression) when $\tau \to \infty$. τ is thus a hyperparameter that generalizes ordinary linear regression so the locally weight linear regression is flexible to how much the user weighs data points that are far from the input data point.

Disadvantage 1: Since locally weighted linear regression requires recomputing the weights for each input data point, this causes the running time to be asymptotically longer than that of ordinary linear regression.

Disadvantage 2: Unlike ordinary linear regression, locally weighted linear regression does not produce an equation that will apply to all future data points. In ordinary linear regression, it is possible to come up with a scenario where given features A and B, a fixed equation $y = w_1 a + w_2 b$ can be produced and evaluated against the test examples to determine the types of (a, b) value (e.g. in a fixed ratio) that have the best predictions y. Once this is known, the equation can be considered optimal specifically for such pairs (a, b) and be used for this specific purpose. Such a possibility does not exist in locally weight linear regression because more than one equation is produced and moreover these equations are specifically trained for only one data point which disallows generalization.