1a) Voluntary bias (individuals choose to respond)

Undercoverage bias (only getting responses from those w/ Twitter)

Convenience Sample (not random)

1b) It was discriminating against women b/c not a lot of current Amazon employees are women, so it may have been biased against women.

Dropping gender may not have been enough b/c of universities, organizations, or opportunities available to women, so it might have "developed" a bias against those.

names

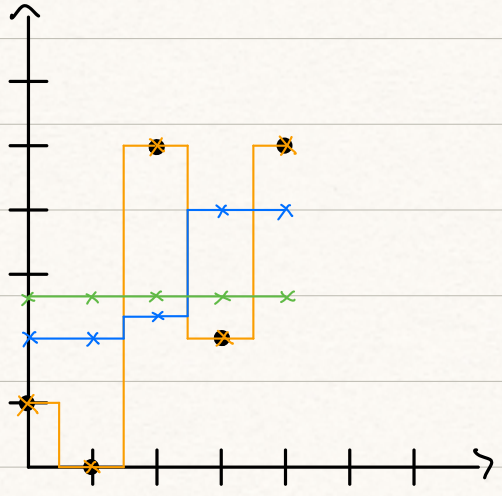
2a) 1-NN: $x=0 \Rightarrow y=1$

$x=1 \Rightarrow y=0$

$x=2 \Rightarrow y=5$

$x=3 \Rightarrow y=2$

$x=4 \Rightarrow y=5$



1-NN

3-NN

5-NN

3-NN: $x=0 \Rightarrow y=\frac{1}{3}(1+0+5)=2$

$x=1 \Rightarrow y=\frac{1}{3}(1+0+5)=2$

$x=2 \Rightarrow y=\frac{1}{3}(0+5+2)=\frac{7}{3}$

$x=3 \Rightarrow y=\frac{1}{3}(5+2+5)=4$

$x=4 \Rightarrow y=\frac{1}{3}(5+2+5)=4$

5-NN: $x=0 \Rightarrow y=\frac{1}{5}(1+0+5+2+5)=2.6$

$x=1 \Rightarrow y=\frac{1}{5}(1+0+5+2+5)=2.6$

$x=2 \Rightarrow y=\frac{1}{5}(1+0+5+2+5)=2.6$

$x=3 \Rightarrow y=\frac{1}{5}(1+0+5+2+5)=2.6$

$x=4 \Rightarrow y=\frac{1}{5}(1+0+5+2+5)=2.6$

2b) K=1 (0.5, 0)(1.5, 0)(2.5, 2)(3.5, 2) MSE = 3.75

K=2 (0.5, 0.5)(1.5, 2.5)(2.5, 3.5)(3.5, 3.5) MSE = 0.25

K=3 (0.5, 2)(1.5, 2)(2.5, 7/3)(3.5, 4) MSE = 1.44

K=4 (0.5, 2)(1.5, 2)(2.5, 3)(3.5, 3) MSE = 0.75

K=5 (0.5, 2.6)(1.5, 2.6)(2.5, 2.6)(3.5, 2.6) MSE = 1.21

Based off MSE, K=2 has the lowest MSE, thus K=2 is the best.

2c) $R^2 = 1 - \frac{\Sigma(\hat{y}_i - y_i)^2}{\Sigma(\bar{y} - y_i)^2}$

When K=1 on the training data, we have $R^2=1$, which means our model is perfect. However, when applying this model to test data sets, it may not be the best, b/c it's "too precise", & overfitted the training data, so it may not perform well for the test data.

3a) $L(\beta_0, \beta_1) = \frac{1}{n}\sum_{i=1}^{n}(y_i - [\beta_0 + \beta_1 x_i])^2$

$$\frac{\partial L}{\partial \beta_0} = \frac{1}{n}\sum_{i=1}^{n} -2(y_i - [\beta_0 + \beta_1 x_i])$$

$$\frac{\partial L}{\partial \beta_1} = \frac{1}{n}\sum_{i=1}^{n} -2x_i(y_i - [\beta_0 + \beta_1 x_i])$$

$$\frac{1}{n}\sum_{i=1}^{n} -2(y_i - [\beta_0 + \beta_1 x_i]) = 0$$

$$\frac{-2}{n}\left[(-n\beta_0) + \sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i\right] = 0$$

$$n\beta_0 = \sum_{i=1}^{n} y_i - \beta_1 x_i$$

$$\beta_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \beta_1 x_i$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$\frac{1}{n}\sum_{i=1}^{n} -2x_i(y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\sum x_i y_i - x_i\beta_0 - \beta_1 x_i^2 = 0$$

$$\sum x_i y_i - \beta_0 \sum x_i = \beta_1 \sum x_i^2$$

$$\sum x_i y_i - (\bar{y} - \beta_1 \bar{x}) \sum x_i = \beta_1 \sum x_i^2$$

$$\sum x_i y_i - \bar{y}\sum x_i + \beta_1 \bar{x} \sum x_i = \beta_1 \sum x_i^2$$

$$\sum x_i y_i - \frac{1}{n}\sum y_i \sum x_i = \beta_1 \sum x_i^2 - \beta_1 \frac{1}{n}\sum x_i \sum x_i$$

$$\sum x_i y_i - \frac{1}{n}(n\bar{y})(n\bar{x}) = \beta_1\left(\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2\right)$$

$$\sum x_i y_i - n\bar{x}\bar{y} = \beta_1\left(\sum x_i^2 - \frac{1}{n}(n\bar{x})^2\right)$$

$$\sum x_i y_i - n\bar{x}\bar{y} = \beta_1\left(\sum x_i^2 - n\bar{x}^2\right)$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = \beta_1\left(\sum (x_i - \bar{x})^2\right)$$

$$\beta_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

3b) MSE: $\frac{1}{n}\sum_{i=1}^{n}(y_i - [\beta_0 + \beta_1 x_i])^2 \Rightarrow L(\beta_0) = \frac{1}{n}\sum_{i=0}^{n}(y_i - \beta_0)^2$

$$\frac{\partial L}{\partial \beta_0} = \frac{1}{n}\sum_{i=0}^{n} 2(y_i - \beta_0)(-1)$$

$$\frac{1}{n}\sum_{i=0}^{n} -2(y_i - \beta_0) = 0$$

$$\sum_{i=0}^{n} y_i - \sum_{i=0}^{n} \beta_0 = 0$$

$$n\bar{y} = n\beta_0 \Rightarrow \beta_0 = \bar{y} \Rightarrow$$ the optimal value for β_0 is $\bar{y}$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \beta_0|$$

$$\frac{\partial f}{\partial \beta_0} = \frac{1}{n}\sum \frac{y_i - \beta_0}{|y_i - \beta_0|}$$

$$= \text{sign}(y_i - \beta_0)$$

The optimal value of β_0 is when there are an equal number of y values greater than β_0 and less than β_0, which is, by definition, the median

4.1.a) $\beta_0 = \bar{y} - \beta_1 \bar{x}$ $\quad \beta_1 = \dfrac{\sum x_i y_i - \bar{y} x_i}{\sum x_i^2 - \bar{x} x_i}$

$\bar{y} = 111.2$ million

$\bar{x} = 1900$

$(1800 \cdot 5M - 111.2M(1800) + (1850 \cdot 23M - 111.2M(1850) + (1900 \cdot 76M - 111.2M(1900) +$

$(1950 \cdot 161M - 111.2M(1950) + (2000 \cdot 291M - 111.2M(2000) = -191160M - 163170M - 66880$

$+ 97110 + 359600 = 35500M$

$(1800^2 - (1800)(1900)) + (1850^2 - (1850)(1900)) + (1900^2 - (1900)(1900))$

$+ (1950^2 - (1950)(1900)) + (2000^2 - (2000)(1900)$

$= -180000 - 92500 + 0 + 97500 + 200000 = 25000$

$\beta_1 = 35500M / 25000 = 1,420,000$

$\beta_0 = 111.2M - (\beta_1)(1900) = -2586.8M$

$\beta_1 = 1,420,000$, $\beta_0 = -2586800000$

4.1.b) $R^2 = 1 - \dfrac{\sum (\hat{y}_i - y_i)^2}{\sum (\bar{y} - y_i)^2}$ $\quad \hat{y} = \beta_0 + \beta_1 x_i$

$(-30.8M - 5M)^2 + (40.2M - 23M)^2 + (111.2M - 76M)^2 + (182.2M - 161M)^2$

$+ (253.2M - 291M)^2 = 4.6948 \times 10^{15}$

$(111.2M - 5M)^2 + (111.2M - 23M)^2 + (111.2M - 76M)^2 + (111.2M - 161M)^2$

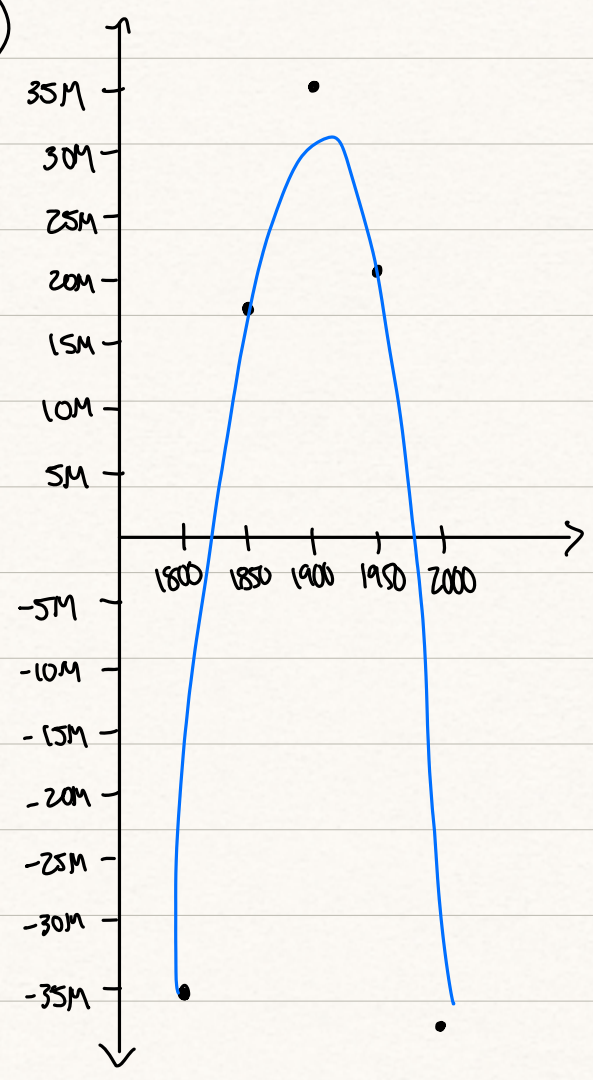
$+ (111.2M - 291M)^2 = 5.51048 \times 10^{16}$

$R^2 = 1 - 0.0852 = 0.915$

We have $R^2 = 0.915$, which is very close to 1, indicating that our model is very good.

(-35.8, 1800) (17.2, 1850) (35.2, 1900)
(21.2, 1950) (-37.8, 2000)

4.1.c)



From the residual graph, I conclude that population growth from 1800-2000 is not necessarily linear.

4.2) While we observe a fairly strong negative correlation between wine consumption & heart disease (as indicated by $R^2 = 71.0\%$), we cannot conclude that drinking more wine reduces the risk of heart disease, b/c correlation does not indicate causation.

4.3.a) (i) $\beta_0 = -36.83$, $\beta_1 = 1.47$

(ii) $\beta_0 = 113.98$, $\beta_1 = -0.681$

4.3.b) (i) $R^2 = 0.658$

(ii) $R^2 = 0.639$

It seems that as the percentage of sunny days increases, less pesticide is used. It also appears that as pesticide use increases, the amount of diseased crops increases. It may appear that the disease rate increases as more pesticide is used, but it could be possible that the increased pesticide use is in response to disease.

4.4) a) $\beta_0 = 5.006$, $\beta_1 = 0.1$

b) $R^2 = 0.243$, This does not show a strong linear relation btw x & y

c) We have p-value: 2.438×10^{-64}. $2.438 \times 10^{-64} < 0.05$, so we reject the null hypothesis.

d) $\beta_1 \pm 2 \times SE(\beta_1)$

Confidence Interval: $[0.0886, 0.111]$

It suggests that β_1 is not meaningfully different from 0

e) The contradiction is that in (c), we reject the hypothesis that $\beta_1 = 0$, but in (d), we conclude that since our $\beta_1 \leq 1$, it is not meaningfully different from zero. This contradiction likely comes from the large range of y values in our dataset. In the future, it may be better to do some data cleaning to ensure better results.

4.5) a) $R^2 = 0.75$, indicating there is a strong linear relationship b/w duration of the last eruption & time until the next eruption.

b) $\hat{y} = 60.383$

prediction interval: $[57.784, 62.983]$

We can be 95% confident that if we observe an eruption that is 3 minutes long, the next eruption will happen in $[57.784, 62.983]$ minutes.

c) Given only 60 minutes, I would be able to make a decision, given that 60 minutes falls within the prediction interval.

5) a) One way to model this would be to have plant type be a categorical variable. X_1 is a numerical variable representing how much sunlight a plant got. X_2 and X_3 are categorical variables for plant type: If we have plant type 1, $X_2 = 0$ & $X_3 = 0$. If we have plant type 2, $X_2 = 1$ & $X_3 = 0$. Else, we have plant type 3, when $X_2 = 1$ and $X_3 = 1$.

Our linear model is $Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_0$.

β_0 represents the average growth for plant type 1 when exposed to no sun, β_1 represents the average growth in all plant types when exposed to sun. β_2 is the diff. in plant growth b/w plant types 1 and 2, and β_3 is the avg. diff in plant growth b/w plant types 1 and 3.

b) To test whether the independent variables have a significant correlation w/ our dependent variable Y, we can calculate t-test values for $\beta_1, \beta_2, \beta_3$. We can then get corresponding p-values. If these p-values are small, then we reject our null hypothesis that there is no relationship b/w Y and the predictor variable.

6) With random sampling, only 5% of data points have $y=0$, so the issue is that the random sample may not accurately represent the dataset. To address this, we can take a stratified sample, where 5% of our sample contains points where $y=0$, and the other 95% contain points where $y=1$.