Exercise 1: Generate simulated that that is similar to the

Exercise 2 (Divide by four rule)**:** Test the divide by for rule on the support for same-sex marriage data set by performing a linear regression on the binned data and comparing the coefficients to the logistic regression output. Clearly explain why this tests the divide by four rule.

Exercise 3:

```
> data = pd.read_csv("https://raw.githubusercontent.com/washingtonpost/data-homicides/master/homicide-da
> data["victim_age"] = pd.to_numeric(data["victim_age"],errors="coerce")
>
> disp_new = []
> disp = data["disposition"].values
> for d in disp:
>   if d == "Closed by arrest":
>     disp_new.append(1)
>   else:
>     disp_new.append(0)
> data["arrest"] = disp_new
> data = data.dropna()
```

Exercise 4 (Water contamination model)**:** Suppose we have data consisting of the longitude (X_1) and latitude (X_2) of water samples along with whether or not the sample tested possitive for a toxin (Y). We believe that their has been a leak and want to identify the region within which there is more than a 1% chance the water is contiminated.

How do we frame this problem as a logistic regression? We need to introduce features so that the contours of equal q (the chance to be contaminated as a function of X_1 and X_2 are ellipsis. Recall the equation is

$$W = a_1X_1 + a_2X_2 + a_3X_1^2 + a_4X_2^2 + a_5X_1X_2 \tag{1}$$

The function $W(X_1, X_2)$ is constant along the curves.

This means that $h(W)$ is constant along these curves, where h is the logistic function. It other words, h will be largest at the center and decay.