1. This dataset contain 15 columns and 32561 rows. It was extracted from the census bureau database at http://www.census.gov/ftp/pub/DES/www/welcome.html. And the reason for creating this dataset is to analyze the relationship between variables such as income, hours worked per week and other variables.
2. Can we predict hours_per_week by relationship and occupation?
3. Hours_per_week, relationship and occupation are included my model since what I want to research is the relationship between hours_per_week and relationship, occupation.
   I try to fit relationship and occupation with hours_per_week in a regression model respectively. So I can know whether there is a statistical significant associations between the two variables. And then generate another regression model with hours_per_week, relationship and occupation together.
4. For the final regression model, we have

$$\beta_0 = 41.67$$
$$\beta_{Armed-Forces} = 2.28$$
$$\beta_{Craft-repai} = 2.34$$
$$\beta_{Exec-managerial} = 5.18$$
$$\beta_{Farming-fishi} = 7.65$$
$$\beta_{Handlers-clea} = 0.38$$
$$\beta_{Machine-op-in} = 1.74$$
$$\beta_{Other-service} = -2.48$$
$$\beta_{Prive-house-} = -4.49$$
$$\beta_{Prof-specia} = 3.26$$
$$\beta_{Protective-se} = 3.26$$
$$\beta_{Sale} = 2.20$$
$$\beta_{Tech-support} = 0.56$$
$$\beta_{Transport-mov} = 4.84$$
$$\beta_{Not-in-family} = -2.44$$
$$\beta_{Other-relativ} = -5.16$$
$$\beta_{Own-child} = -9.44$$
$$\beta_{Unmarried} = -3.58$$
$$\beta_{Wife} = -5.75$$

   - The intercept is the estimated expected hours_per_week occupation is Adm-clerical and relationship is Husband.
   - The coefficient for Armed-Forces is the estimated expected difference in hours_per_week when occupation change from Adm-clerical to Armed-Forces, holding sex constant.
   - The coefficient for Craft-repair is the estimated expected difference in hours_per_week when occupation change from Adm-clerical to Craft-repair, holding sex constant. But its p-value is so large that we shouldn't apply this in this model.
   - The coefficient for Exec-managerial is the estimated expected difference in hours_per_week when occupation change from Adm-clerical to Exec-managerial, holding sex constant.
   - The coefficient for Farming-fishing is the estimated expected difference in hours_per_week when occupation change from Adm-clerical to Farming-fishing, holding sex constant.
   - The coefficient for Handlers-cleaners is the estimated expected difference in hours_per_week when occupation change from Adm-clerical to Handlers-cleaners, holding sex constant. But its p-value is so large that we shouldn't apply this in this model.
   - The coefficient for Machine-op-inspect is the estimated expected difference in hours_per_week when occupation change from Adm-clerical to Machine-op-inspect, holding sex constant.
   - The coefficient for Other-service is the estimated expected difference in hours_per_week when occupation change from Adm-clerical to Other-service, holding sex constant.
   - The coefficient for Prive-house-serv is the estimated expected difference in hours_per_week when occupation change from Adm-clerical to Prive-house-serv, holding sex constant.

- The coefficient for Prof-specialty is the estimated expected difference in hours_per_week when occupation change from Adm-clerical to Prof-specialty, holding sex constant.
- The coefficient for Protective-serv is the estimated expected difference in hours_per_week when occupation change from Adm-clerical to Protetive-serve, holding sex constant.
- The coefficient for Sales is the estimated expected difference in hours_per_week when occupation change from Adm-clerical to Sale, holding sex constant. But its p-value is so large that we shouldn't apply this in this model.
- The coefficient for Tech-support is the estimated expected difference in hours_per_week when occupation change from Adm-clerical to Tech-support, holding sex constant.
- The coefficient for Transport-moving is the estimated expected difference in hours_per_week when occupation change from Adm-clerical to Transport-moving, holding sex constant.
- The coefficient for Not-in-family is the estimated expected difference in hours_per_week when relationship change from Husband to Not-in-family, holding occpation constant.
- The coefficient for Other-relative is the estimated expected difference in hours_per_week when relationship change from Husband to Other-relative, holding occpation constant.
- The coefficient for Own-child is the estimated expected difference in hours_per_week when relationship change from Husband to Own-child, holding occpation constant.
- The coefficient for Unmarried is the estimated expected difference in hours_per_week when relationship change from Husband to Unmarried, holding occpation constant.
- The coefficient for Wife is the estimated expected difference in hours_per_week when relationship change from Husband to Wife, holding occpation constant.

$$\hat{\beta}_1 = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} = \frac{\sum(xy - x\bar{y} - \bar{x}y + \bar{x}\bar{y})}{\sum(x^2 - 2\bar{x}x + \bar{x}^2)} = \frac{\sum xy - \bar{y}\sum x - \bar{x}\sum y + n\bar{x}\bar{y}}{\sum x^2 - 2\bar{x}\sum x + n\bar{x}^2}$$

$$= \frac{\sum xy - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y}}{\sum x^2 - 2n\bar{x}^2 + n\bar{x}^2} = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2} = \frac{\frac{1}{n}\sum xy - \bar{x}\bar{y}}{\frac{1}{n}\sum x^2 - \bar{x}^2} = \frac{E(XY) - E(X)E(Y)}{E(X^2) - E^2(X)}$$

$$= \frac{E(XY) - E(X)E(Y)}{\sigma_X^2}$$

$$corr(X,Y) = \frac{E(XY) - E(X)E(Y)}{\sigma_X \sigma_Y} = \frac{E(XY) - \dot{E}(X)E(Y)}{\sigma_X^2} \cdot \frac{\sigma_X}{\sigma_Y} = \hat{\beta}_1 \cdot \frac{\sigma_X}{\sigma_Y}$$

So, $\frac{corr(X,Y)}{\hat{\beta}_1} = \frac{\sigma_X}{\sigma_Y}$

The ratio of the covariance and slope is equal to the ratio of the variance of X and the variance of Y