1. Filtering Data

   - Dataset contains employees whose ages are at least 18 year-old.
   - Filter out the unrelated columns.
   - Filter out the rows where class is either 'Without pay' nor 'Never worked.'
   - Filter out the rows where employment_commitment is either 'Not in labor force', 'Unemployed part- time', or 'Unemployed full-time.'
   - Remove NA values

age	gender	education	class	race	employment_commitment	employment_stat	wage_per_hour	working_week_per_year	industry_code_main	citizenship	income_above_limit
21	Male	12th grade no diploma	Federal government	Black	Children or Armed Forces	0	500	15	Hospital services	Native	Below limit
45	Male	Bachelors degree(BA AB BS)	Private	Asian or Pacific Islander	Children or Armed Forces	0	825	52	Retail trade	Foreign born- Not a citizen of U S	Below limit
53	Male	High school graduate	Private	White	Full-time schedules	0	0	52	Retail trade	Native	Below limit
22	Female	High school graduate	Private	White	Full-time schedules	0	0	52	Finance insurance and real estate	Native	Below limit
22	Female	11th grade	Private	Black	Full-time schedules	0	0	48	Manufacturing-nondurable goods	Native	Below limit
30	Male	High school graduate	Local government	White	Full-time schedules	0	0	39	Transportation	Native	Below limit

2. Change all non-numeric data into numeric data

age	gender	education	class	race	employment_commitment	employment_stat	wage_per_hour	working_week_per_year	citizenship	income_above_limit	industry
21	1	11	0	1	0	0	500	15	1	1	1
45	1	16	1	2	0	0	825	52	0	1	2
53	1	12	1	3	2	0	0	52	1	1	2
22	0	12	1	3	2	0	0	52	1	1	3
22	0	11	1	1	2	0	0	48	1	1	4
30	1	12	2	3	2	0	0	39	1	1	5
43	1	12	3	3	0	2	0	52	1	1	6
40	0	11	1	3	0	0	0	52	1	1	7
47	1	16	2	3	0	0	0	37	1	1	8
22	1	11	1	1	2	0	600	32	1	1	9

3. Datasets

   Create 3 datasets to build the models

   - dataset1 contains the zero values in wage_per_hour.
   - dataset2: Replace the zero values in wage_per_hour with the average wage_per_hour for each gender.

gender	avg_wage_per_hour
0	114.8211
1	118.4655

- dataset3: Replace the zero values in wage_per_hour with the average wage_per_hour for each gender, grouped by above_limit_income

income_above_limit	gender	avg_wage_per_hour
0	0	116.13025
0	1	125.56560
1	0	91.76870
1	1	86.85845

4. Tree model

4.1. Use dataset1 to build a tree model

Summary of the model

```
Classification tree:
tree(formula = income_above_limit ~ ., data = train1)
Variables actually used in tree construction:
[1] "education" "gender"    "age"
Number of terminal nodes:  6
Residual mean deviance:  0.6232 = 48580 / 77960
Misclassification error rate: 0.1225 = 9554 / 77970
```

Confusion Matrix

```
       Predict
Actual     0     1
     0 17077     0
     1  2380     0
```

Accuracy = 0.877679

Precision = 0

4.2. Use dataset2 to build a tree model

Summary of the model

```
Classification tree:
tree(formula = income_above_limit ~ ., data = train2)
Variables actually used in tree construction:
[1] "education" "gender"    "age"
Number of terminal nodes:  6
Residual mean deviance:  0.6232 = 48580 / 77960
Misclassification error rate: 0.1225 = 9554 / 77970
```

Confusion Matrix

```
       Predict
Actual     0     1
     0 17077     0
     1  2380     0
```

Accuracy = 0.877679

Precision = 0

4.3. Use dataset3 to build a tree model

Summary of the model

```
Classification tree:
tree(formula = income_above_limit ~ ., data = train3)
Variables actually used in tree construction:
[1] "wage_per_hour"
Number of terminal nodes:  3
Residual mean deviance:  0.04749 = 3703 / 77970
Misclassification error rate: 0.005964 = 465 / 77970
```

Confusion Matrix

```
       Predict
Actual     0     1
     0 17077     0
     1   106  2274
```

Accuracy = 0.9945521

Precision = 0.9554622