## Data Structure

'data.frame': 19602 obs. of 15 variables:

$ age : int 39 50 38 53 28 37 49 52 31 42 ...

$ workclass : Factor w/ 9 levels " ?"," Federal-gov",..: 8 7 5 5 5 5 5 7 5 5 ...

$ fnlwgt : int 77516 83311 215646 234721 338409 284582 160187 209642 45781 159449 ...

$ education : Factor w/ 16 levels " 10th"," 11th",..: 10 10 12 2 10 13 7 12 13 10 ...

$ education.num : int 13 13 9 7 13 14 5 9 14 13 ...

$ marital.status: Factor w/ 7 levels " Divorced"," Married-AF-spouse",..: 5 3 1 3 3 3 4 3 5 3 ...

$ occupation : Factor w/ 15 levels " ?"," Adm-clerical",..: 2 5 7 7 11 5 9 5 11 5 ...

$ relationship : Factor w/ 6 levels " Husband"," Not-in-family",..: 2 1 2 1 6 6 2 1 2 1 ...

$ race : Factor w/ 5 levels " Amer-Indian-Eskimo",..: 5 5 5 3 3 5 3 5 5 5 ...

$ sex : Factor w/ 3 levels " Fema"," Female",..: 3 3 3 3 2 2 2 3 2 3 ...

$ capital.gain : int 2174 0 0 0 0 0 0 0 14084 5178 ...

$ capital.loss : int 0 0 0 0 0 0 0 0 0 0 ...

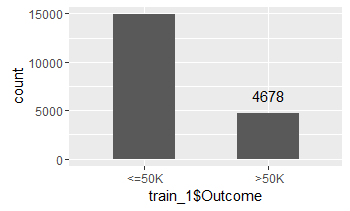
$ hours.per.week: int 40 13 40 40 40 40 16 45 50 40 ...

$ native.country: Factor w/ 42 levels ""," ?"," Cambodia",..: 40 40 40 40 7 40 24 40 40 40 ...

$ Outcome : Factor w/ 3 levels ""," <=50K"," >50K": 2 2 2 2 2 2 2 3 3 3 ...

## Check for data imbalance

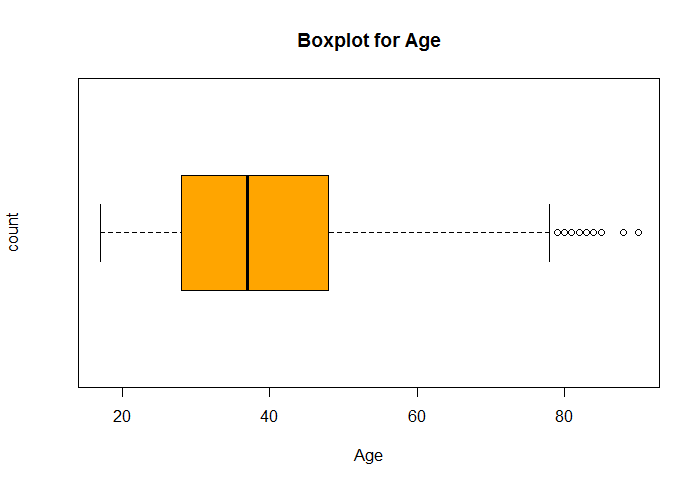
* 1 record with outcome variable as blank in the train data set
* Ignoring that record.
* Train\_1 created with 19601 records.



* + 14923 Vs 4678 – Slightly imbalanced

## Analyze each column

### Age





* Slightly right skewed
* Few outliers on the right side
* 0 missing values

### Workclass

* > NROW(train\_1[which(trim(train\_1$workclass)=="?"),])
* [1] 1096
* 1096 records with unknown/missing values
* More private employees



### FnlWgt



* Right Skewed distribution
* Outliers exist
* 0 missing values

