Numerical bucketing or binning

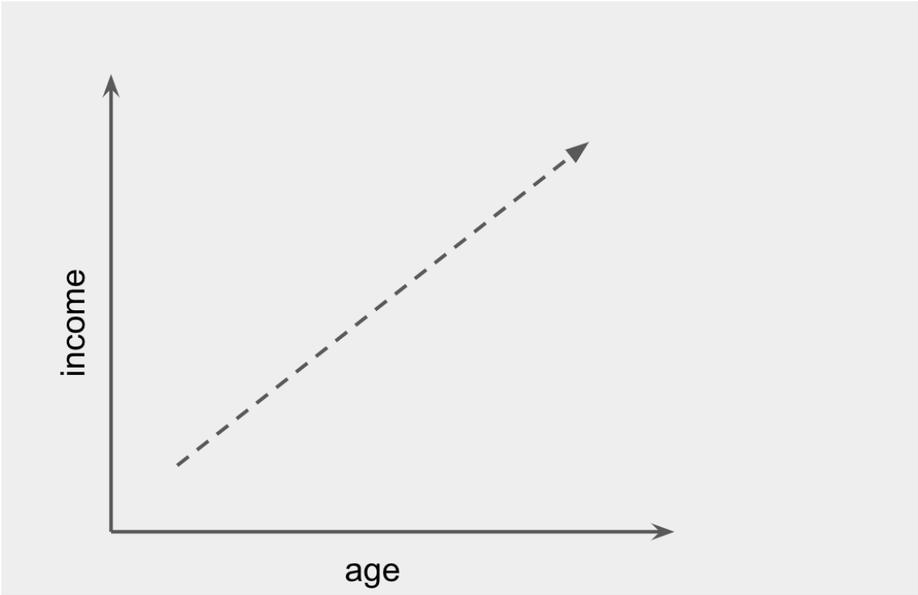
FEATURE ENGINEERING IN R

Jose Hernandez
Data Scientist, University of Washington

Income age relationship

```
adult_incomes %>%
    select(age) %>%
    glimpse()
```

```
int [1:32561] 39 50 38 53 28 37 49 52 31 42 ...
```



Income age relationship

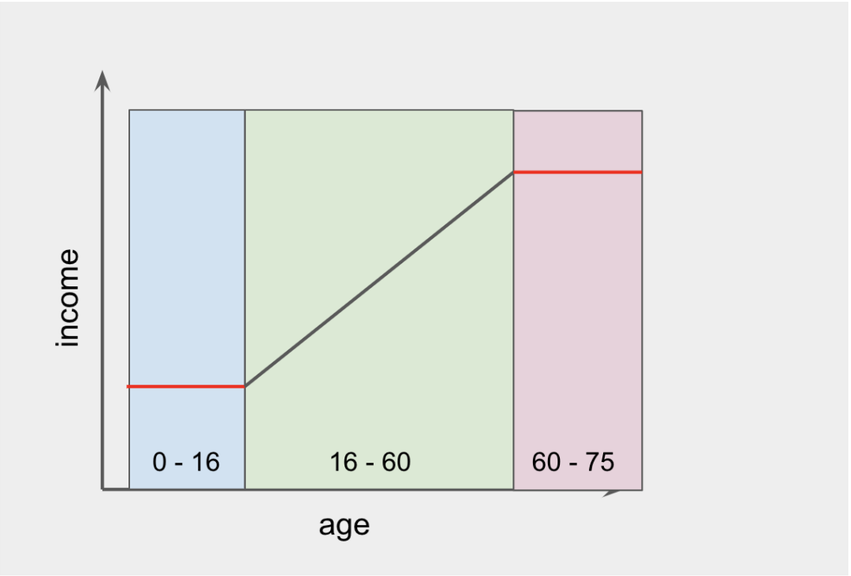
```
adult_incomes %>%
    select(age) %>%
    glimpse()
```

```
int [1:32561] 39 50 38 53 28 37 49 52 31 42 ...
```

Income age relationship

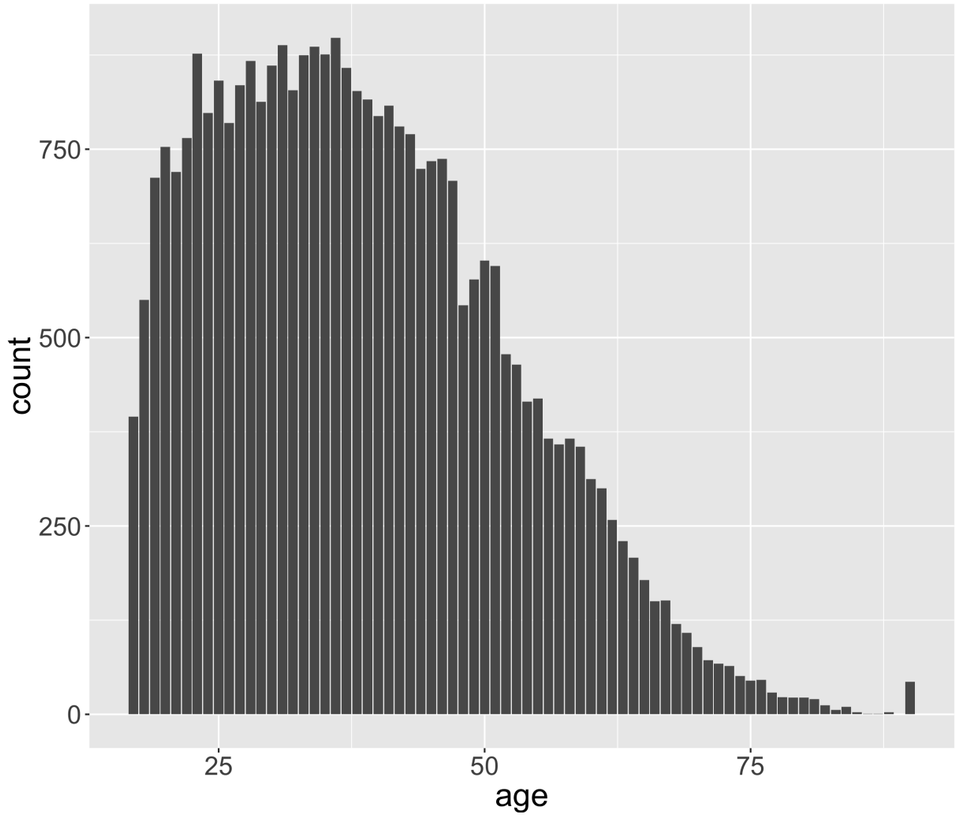
```
adult_incomes %>%
    select(age) %>%
    glimpse()
```

```
int [1:32561] 39 50 38 53 28 37 49 52 31 42 ...
```



Income variable distribution

```
ggplot(adult_incomes, aes(x = age)) + geom_histogram(stat = "count")
```



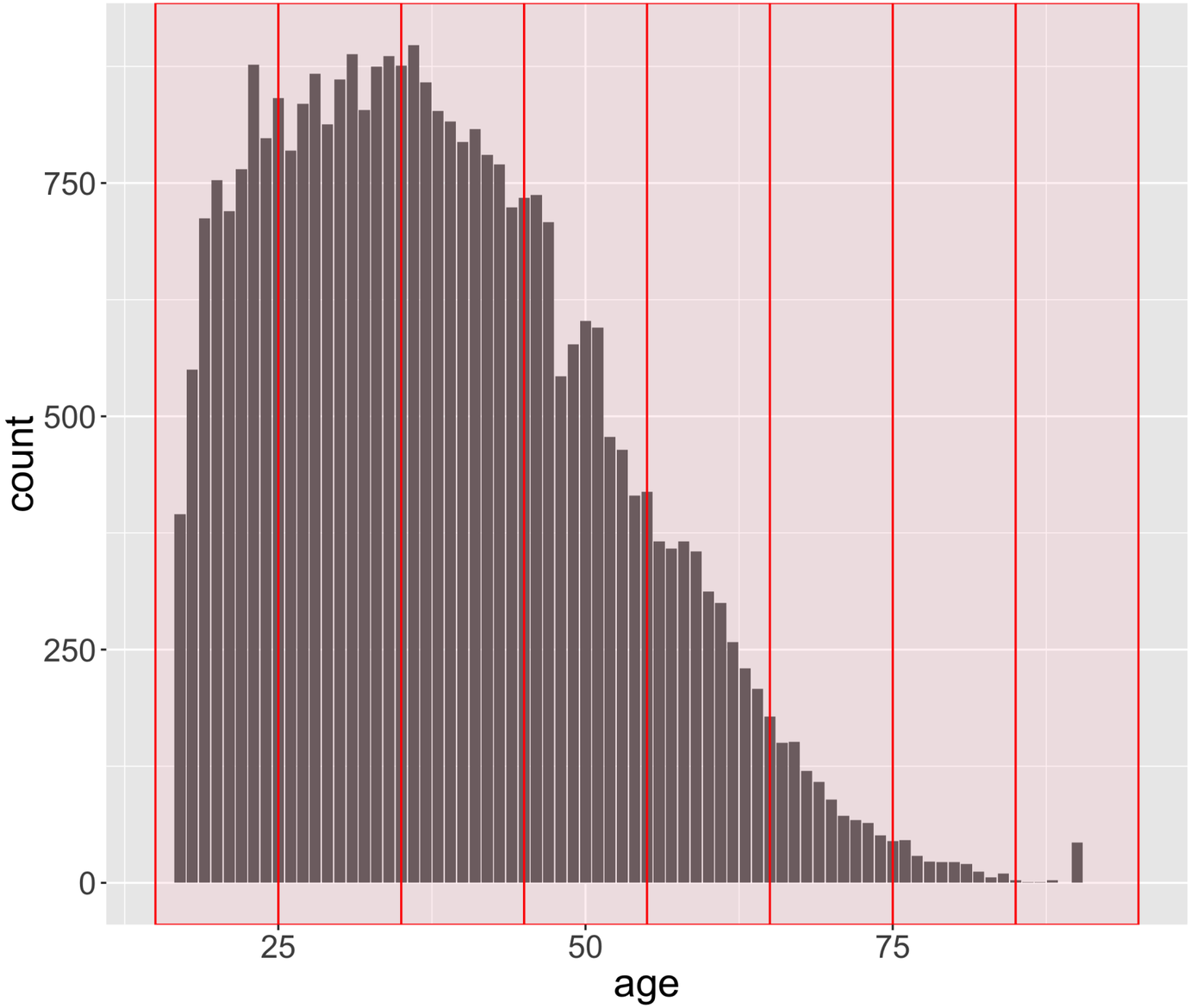
Enter numerical bucketing

```
adult_incomes %>%
    select(age) %>%
    summary()
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  17.00   28.00   37.00   38.58   48.00   90.00
```

```
seq(15, 95, by = 10)
```

```
[1] 15 25 35 45 55 65 75 85 95
```


750
500
250
0
count
25
50
75
age

```
 adult_incomes %>%
   mutate(age_cat = cut(age, breaks = seq(15, 95, by = 10))) %>%
   select(age_cat) %>% table()
```

```
(15,25] (25,35] (35,45] (45,55] (55,65] (65,75] (75,85] (85,95]
   6411    8514    8009    5538    2931     917     193      48
```

```
adult_incomes %>%
   mutate(age_cat=cut(age, breaks = seq(15, 95, by = 10))) %>%
   select(age, age_cat) %>% data.frame() %>% head()
```

```
  age age_cat
1  39 (35,45]
2  50 (45,55]
3  38 (35,45]
4  53 (45,55]
5  28 (25,35]
6  37 (35,45]
```

```
dmy_data <- model.matrix(income ~ age_cat - 1, data = out_data)
```

```
head(dmy_data)
```

```
  age_cat(15,25] age_cat(25,35] age_cat(35,45] age_cat(45,55]
1              0              0              1              0
2              0              0              0              1
3              0              0              1              0
4              0              0              0              1
5              0              1              0              0
6              0              0              1              0
  age_cat(55,65] age_cat(65,75] age_cat(75,85] age_cat(85,95]
1              0              0              0              0
2              0              0              0              0
3              0              0              0              0
4              0              0              0              0
5              0              0              0              0
6              0              0              0              0
```

It's your turn!

FEATURE ENGINEERING IN R

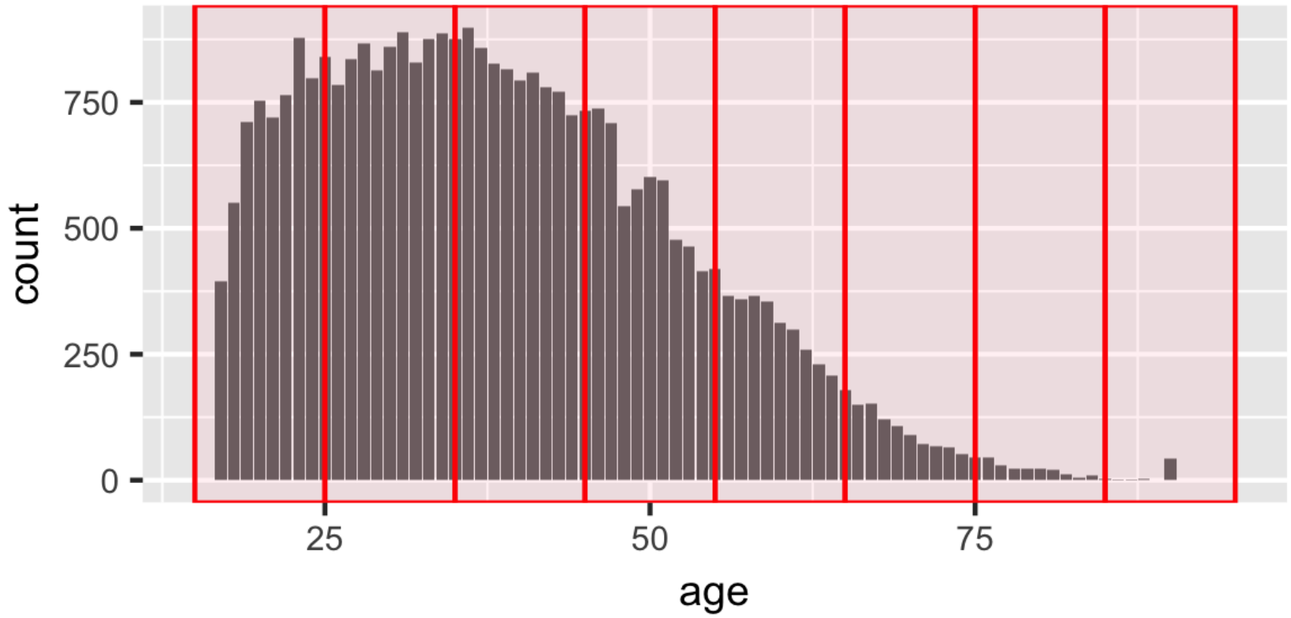
Binning numerical data using quantiles

FEATURE ENGINEERING IN R

Jose Hernandez
Data Scientist, University of Washington

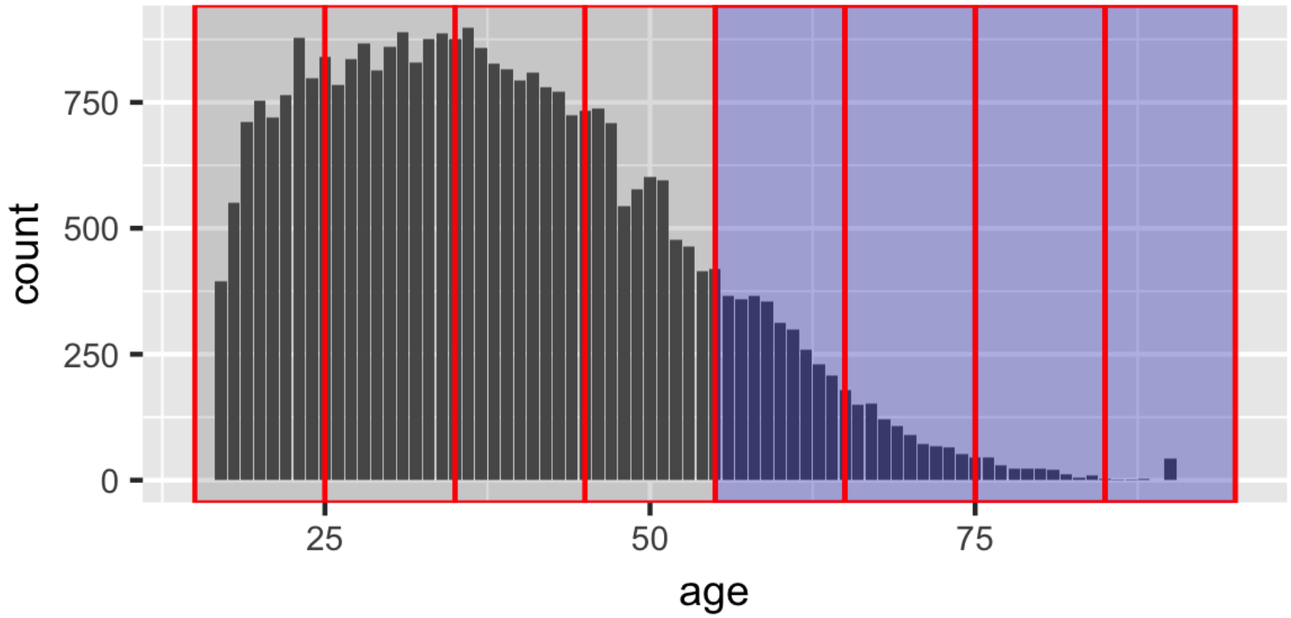
```
adult_incomes %>%
   mutate(age_cat = cut(age, breaks = seq(15, 95, by = 10))) %>%
   select(age_cat) %>%
   table()
```

```
(15,25] (25,35] (35,45] (45,55] (55,65] (65,75] (75,85] (85,95]
   6411    8514    8009    5538    2931     917     193      48
```



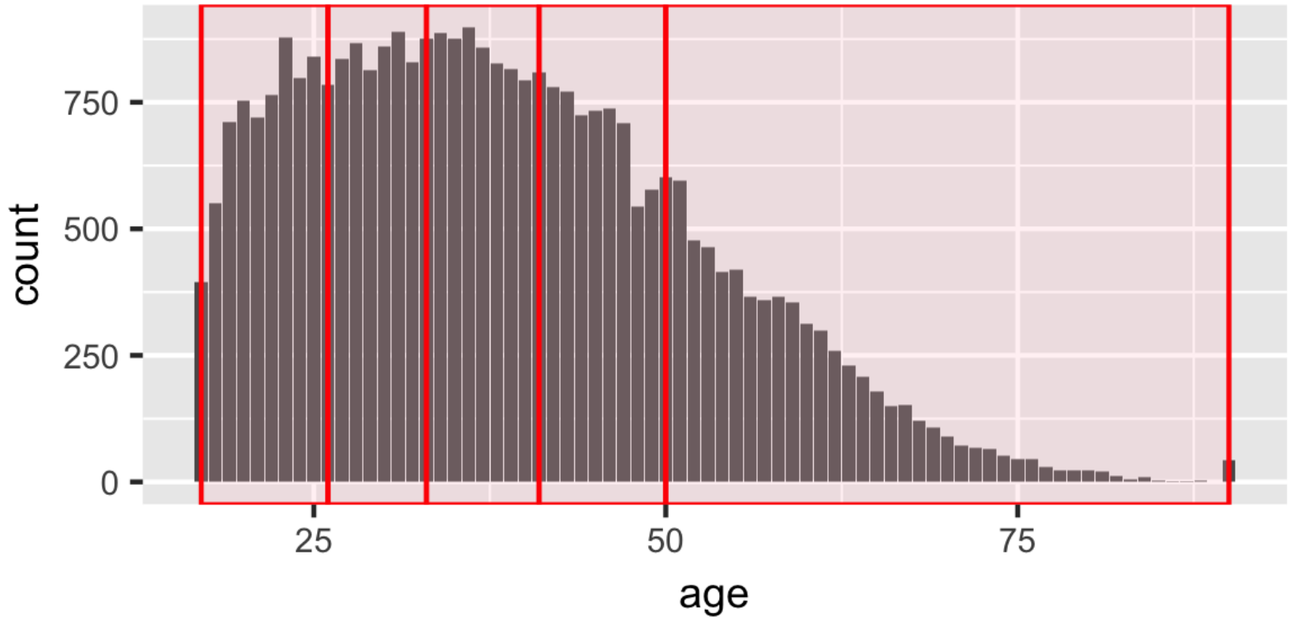
```
adult_incomes %>%
   mutate(age_cat = cut(age, breaks = seq(15, 95, by = 10))) %>%
   select(age_cat) %>%
   table()
```

```
(15,25] (25,35] (35,45] (45,55] (55,65] (65,75] (75,85] (85,95]
   6411    8514    8009    5538    2931     917     193      48
```



```
# Quantile bucketing
adult_incomes %>%
  mutate(age_q = ntile(age, 5)) %>%
  select(age_q) %>%
  table()
```

```
   1    2    3    4    5
6513 6512 6512 6512 6512
```



```
# Variable age ranges
adult_incomes %>%
   mutate(age_q = ntile(age, 5)) %>%
   group_by(age_q, age) %>%
   summarize(n = n()) %>%
   group_by(age_q) %>%
   summarize(total = sum(n),
             min_age = min(age),
             max_age = max(age))
```

```
# A tibble: 5 x 4
  age_q total min_age max_age
  <int> <int>   <dbl>   <dbl>
1     1  6513      17      26
2     2  6512      26      33
3     3  6512      33      41
4     4  6512      41      50
5     5  6512      50      90
```

Converting to actual features

```
dmy_data <- model.matrix(~ age_q - 1, data = adult_incomes)
```

```
head(dmy_data)
```

```
  age_q1 age_q2 age_q3 age_q4 age_q5
1      0      0      1      0      0
2      0      0      0      1      0
3      0      0      1      0      0
4      0      0      0      0      1
5      0      1      0      0      0
6      0      0      1      0      0
```

It's your turn!

FEATURE ENGINEERING IN R

Date and time feature extraction

FEATURE ENGINEERING IN R

Jose Hernandez
Data Scientist, University of Washington

Looking at online retail sales

```
glimpse(online_retail)
```

```
Observations: 100,000
Variables: 8
$ InvoiceNo   <chr> "536365", "536365", "536365", "536365", "536365", ...
$ StockCode   <chr> "85123A", "71053", "84406B", "84029G", "84029E", "...
$ Description <chr> "WHITE HANGING HEART T-LIGHT HOLDER", "WHITE METAL...
$ Quantity    <int> 6, 6, 8, 6, 6, 2, 6, 6, 6, 32, 6, 6, 8, 6, 6, 3, 2...
$ InvoiceDate <chr> "12/1/10 8:26", "12/1/10 8:26", "12/1/10 8:26", "1...
$ UnitPrice   <dbl> 2.55, 3.39, 2.75, 3.39, 3.39, 7.65, 4.25, 1.85, 1....
$ CustomerID  <int> 17850, 17850, 17850, 17850, 17850, 17850, 17850, 1...
$ Country     <chr> "United Kingdom", "United Kingdom", "United Kingdo...
```

```
online_retail %>%
    select(InvoiceDate) %>%
    glimpse()
```

```
chr [1:100000] "12/1/10 8:26" "12/1/10 8:26" "12/1/10 8:26"
```

```
as.Date("12/1/10", format = "%m/%d/%y")
```

```
[1] "2010-12-01"
```

```
library(lubridate)
mdy_hm("12/1/10 8:30")
```

```
[1] "2010-12-01 08:30:00 UTC"
```

- **ymd_hm()** = *"Year, month, day, hour, minutes"*
- **ymd_hms()** = *"Year, month, day, hour, minutes, seconds"*

```
online_retail %>%
  mutate(InvoiceDate = mdy_hm(InvoiceDate)) %>%
  select(InvoiceDate) %>%
  glimpse()
```

```
$ InvoiceDate <dttm> 2010-12-01 08:26:00, 2010-12-01 08:26:00
```

```
wday("12/1/10 8:30")
```

```
[1] 3
```

```
wday("12/1/10 8:30", label = TRUE)
```

```
[1] Tue
Levels: Sun < Mon < Tue < Wed < Thu < Fri < Sat
```

```
# Extracting day of the week
online_retail <- online_retail %>%
  mutate(dow = lubridate::wday(InvoiceDate, label = TRUE))

glimpse(online_retail)
```

```
Observations: 100,000
Variables: 10
$ InvoiceNo    <chr> "536365", "536365", "536365", "536365", "53...
$ StockCode    <chr> "85123A", "71053", "84406B", "84029G", "840...
$ Description  <chr> "WHITE HANGING HEART T-LIGHT HOLDER", "WHIT...
$ Quantity     <int> 6, 6, 8, 6, 6, 2, 6, 6, 6, 32, 6, 6, 8, 6, ...
$ InvoiceDate  <dttm> 2010-12-01 08:26:00, 2010-12-01 08:26:00, ...
$ UnitPrice    <dbl> 2.55, 3.39, 2.75, 3.39, 3.39, 7.65, 4.25, 1...
$ CustomerID   <int> 17850, 17850, 17850, 17850, 17850, 17850, 1...
$ Country      <chr> "United Kingdom", "United Kingdom", "United...
$ InvoiceDate2 <dttm> 2010-12-01 08:26:00, 2010-12-01 08:26:00, ...
$ dow          <ord> Wed, Wed, Wed, Wed, Wed, Wed, Wed, Wed, Wed...
```

```
hour("12/1/10 8:30")
```

```
[1] 8
```

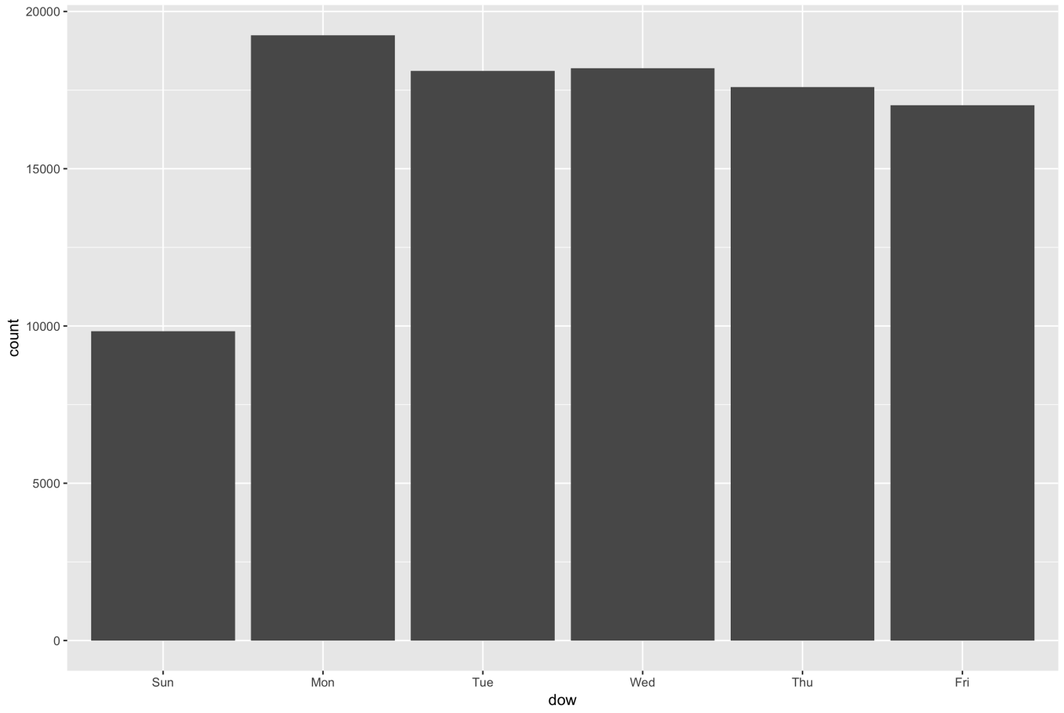
```
online_retail <- online_retail %>%
  mutate(hod = lubridate::hour(InvoiceDate))

glimpse(online_retail)
```

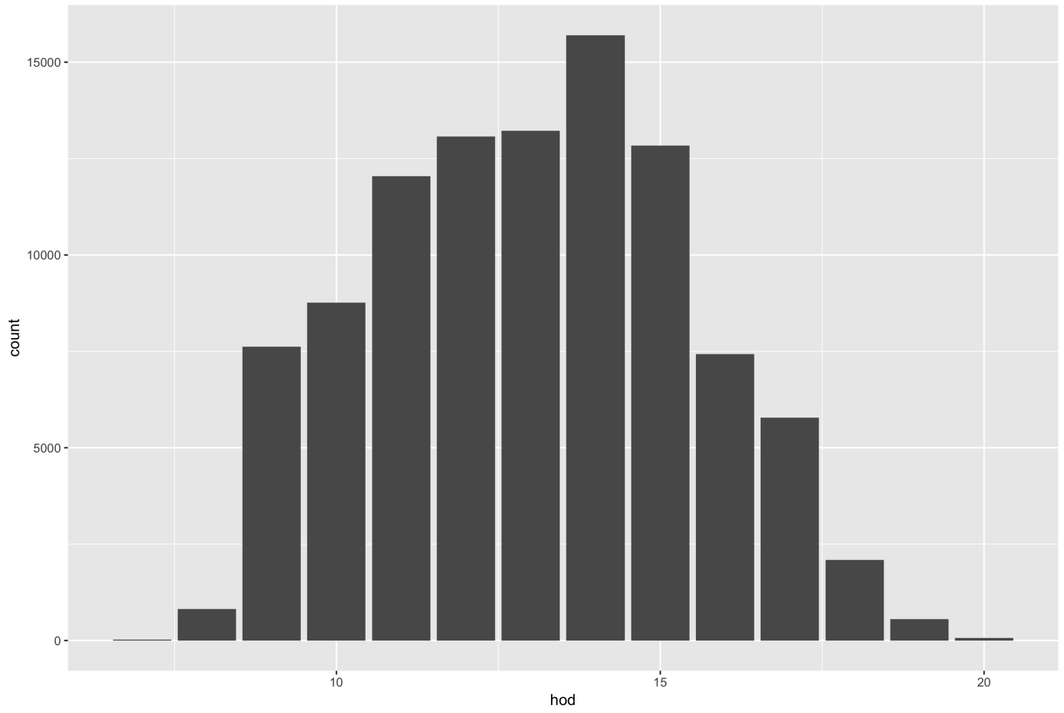
```
Observations: 100,000
Variables: 11
$ InvoiceNo    <chr> "536365", "536365", "536365", "536365", "53...
$ StockCode    <chr> "85123A", "71053", "84406B", "84029G", "840...
$ Description  <chr> "WHITE HANGING HEART T-LIGHT HOLDER", "WHIT...
$ Quantity     <int> 6, 6, 8, 6, 6, 2, 6, 6, 6, 32, 6, 6, 8, 6, ...
$ InvoiceDate  <dttm> 2010-12-01 08:26:00, 2010-12-01 08:26:00, ...
$ UnitPrice    <dbl> 2.55, 3.39, 2.75, 3.39, 3.39, 7.65, 4.25, 1...
$ CustomerID   <int> 17850, 17850, 17850, 17850, 17850, 17850, 1...
$ Country      <chr> "United Kingdom", "United Kingdom", "United...
$ InvoiceDate2 <dttm> 2010-12-01 08:26:00, 2010-12-01 08:26:00, ...
$ dow          <dbl> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4...
$ hod          <int> 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8...
```

Visualizing new features

```
ggplot(online_retail, aes(x = dow)) +
  geom_histogram(stat = "count")
```




```
ggplot(online_retail,aes(x=hod)) +
  geom_histogram(stat = "count")
```



Let's practice!

FEATURE ENGINEERING IN R