《R数据科学》 第一次作业

李游游

2023-10-24

第一题：探索nycflights13数据集

1. 从flights数据中找出到达时间延误2小时或者更多的所有航班，并将生成的新数据，保存为 flight_arr2hr。

如下面表格所示，展示出前10条示例数据

```
## # A tibble: 10,200 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
##  1  2013     1     1      811            630       101     1047            830
##  2  2013     1     1      848           1835       853     1001           1950
##  3  2013     1     1      957            733       144     1056            853
##  4  2013     1     1     1114            900       134     1447           1222
##  5  2013     1     1     1505           1310       115     1638           1431
##  6  2013     1     1     1525           1340       105     1831           1626
##  7  2013     1     1     1549           1445        64     1912           1656
##  8  2013     1     1     1558           1359       119     1718           1515
##  9  2013     1     1     1732           1630        62     2028           1825
## 10  2013     1     1     1803           1620       103     2008           1750
## # i 10,190 more rows
## # i 11 more variables: arr_delay <dbl>, carrier <chr>, flight <int>,
## #   tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>,
## #   hour <dbl>, minute <dbl>, time_hour <dttm>
```

2. 将生成的flight_arr2hr数据集根据目的地(dest)进行分组，统计出抵达每个目的地的航班数量，筛选出抵达航班数量前十名
   top10_dest。

如下面表格所示

```
## # A tibble: 10 x 2
##    dest  count
##    <chr> <int>
##  1 ATL     582
##  2 BOS     355
##  3 CLT     367
##  4 DTW     277
##  5 FLL     384
##  6 IAD     269
##  7 LAX     318
##  8 MCO     392
##  9 ORD     578
## 10 SFO     413
```

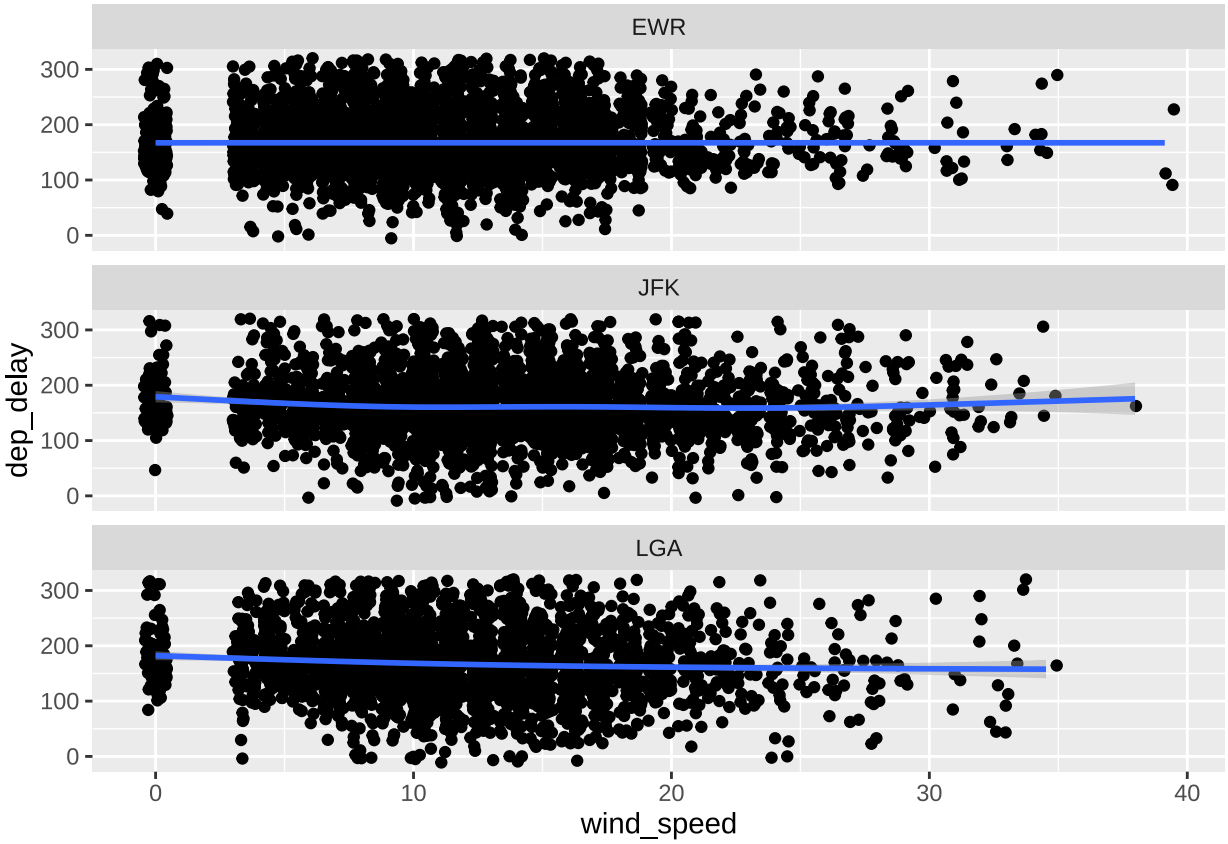
3. 从 weather 表中挑选出以下变量:year, month, day, hour, origin, humid, wind_speed，并将其与 flight_arr2hr 表根据共同变量进行左连接, 生成的新数据保存为 flight_weather

数据如下表所示

```
## # A tibble: 10,200 x 21
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
##  1  2013     1     1      811            630       101     1047            830
##  2  2013     1     1      848           1835       853     1001           1950
##  3  2013     1     1      957            733       144     1056            853
##  4  2013     1     1     1114            900       134     1447           1222
##  5  2013     1     1     1505           1310       115     1638           1431
##  6  2013     1     1     1525           1340       105     1831           1626
##  7  2013     1     1     1549           1445        64     1912           1656
##  8  2013     1     1     1558           1359       119     1718           1515
##  9  2013     1     1     1732           1630        62     2028           1825
## 10  2013     1     1     1803           1620       103     2008           1750
## # i 10,190 more rows
## # i 13 more variables: arr_delay <dbl>, carrier <chr>, flight <int>,
## #   tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>,
## #   hour <dbl>, minute <dbl>, time_hour <dttm>, humid <dbl>, wind_speed <dbl>
```

4. 基于 flight_weather 数据集，根据不同出发地(origin)在平行的三个图中画出风速wind_speed(x 轴)和出发延误时间dep_delay

如下图所示(去掉dep_delay高处离群点)



5. flights中每家航空公司在2013年有多少班次的航班被取消了?提示:依据dep_time来判断某班次航班是否被取消

每家航空公司在2013年的航班取消数据如下表，共有8255个班次被取消

```
## # A tibble: 15 x 2
##    carrier cancelled_count
##    <chr>             <int>
##  1 9E                 1044
##  2 AA                  636
##  3 AS                    2
##  4 B6                  466
##  5 DL                  349
##  6 EV                 2817
##  7 F9                    3
##  8 FL                   73
##  9 MQ                 1234
## 10 OO                    3
## 11 UA                  686
## 12 US                  663
## 13 VX                   31
## 14 WN                  192
## 15 YV                   56
```

6. 找出flights中每一家航空公司的航班最常去的目的地机场，以及flights中每家航空公司飞往最常去的目的地机场的航班数

结果如下表所示

```
## # A tibble: 16 x 3
## # Groups:   carrier [16]
##    carrier dest  flights_count
##    <chr>   <chr>         <int>
##  1 9E      CVG            1559
##  2 AA      DFW            7257
##  3 AS      SEA             714
##  4 B6      FLL            6563
##  5 DL      ATL           10571
##  6 EV      IAD            4048
##  7 F9      DEN             685
##  8 FL      ATL            2337
##  9 HA      HNL             342
## 10 MQ      RDU            4794
## 11 OO      CLE              24
## 12 UA      ORD            6984
## 13 US      CLT            8632
## 14 VX      LAX            2580
## 15 WN      MDW            4113
## 16 YV      IAD             311
```

第二题：数据连接及画图

1. 请将数据 hw1_a 和 hw1_b 分别读入 R，查看数据并指出各个变量的形式，最小值，最大值，中值，均值，标准差

```
## [1] "hw1_a details:"

## [1] "Variable  ID"
## [1] "class:  numeric"
## [1] "min:  1"
## [1] "max:  200"
## [1] "median:  98"
## [1] "mean:  98.8148148148148"
## [1] "sd:  57.3208828363717"

## [1] "Variable  Age"
## [1] "class:  numeric"
## [1] "min:  20.1895762110035"
## [1] "max:  55.7240626717859"
## [1] "median:  33.2566347855657"
## [1] "mean:  34.9624273473237"
## [1] "sd:  8.2546650241916"

## [1] "Variable  Years_at_Employer"
## [1] "class:  numeric"
## [1] "min:  0.143477291896765"
## [1] "max:  31.646028794314"
## [1] "median:  7.62926301226487"
## [1] "mean:  8.90279919988887"
## [1] "sd:  6.83545102331253"

## [1] "Variable  Years_at_Address"
## [1] "class:  numeric"
## [1] "min:  0.00510755169428858"
## [1] "max:  3.69614788285423"
## [1] "median:  0.620697145475018"
## [1] "mean:  0.784321963928919"
## [1] "sd:  0.636265765325646"

## [1] "Variable  Income"
## [1] "class:  numeric"
## [1] "min:  11522.1012336204"
## [1] "max:  451319.666749404"
## [1] "median:  34375.0846904312"
## [1] "mean:  49626.0635495"
## [1] "sd:  49034.3112851568"

## [1] "hw1_b details:"

## [1] "Variable  ID"
## [1] "class:  numeric"
## [1] "min:  1"
## [1] "max:  200"
## [1] "median:  100"
## [1] "mean:  101.460317460317"
## [1] "sd:  57.9657342095472"
```

```
## [1] "Variable  Credit_Card_Debt"
## [1] "class:  numeric"
## [1] "min:  -32050.3773558026"
## [1] "max:  34.1638172378165"
## [1] "median:  -1833.33189317958"
## [1] "mean:  -3287.11075418808"
## [1] "sd:  3972.94897128679"

## [1] "Variable  Automobile_Debt"
## [1] "class:  numeric"
## [1] "min:  -55418.5675346635"
## [1] "max:  1747.23519016687"
## [1] "median:  -3964.40838617404"
## [1] "mean:  -6429.26464822862"
## [1] "sd:  7569.76955542074"

## [1] "Variable  Is_Default"
## [1] "class:  numeric"
## [1] "min:  0"
## [1] "max:  1"
## [1] "median:  0"
## [1] "mean:  0.248677248677249"
## [1] "sd:  0.433394379094478"
```

2. 结合上课我们所学的几种数据 join 的形式，尝试将两个数据集进行合并。对于每种数据合并的方式，请说明 key, 并且报告合并后的数据样本总行数

课堂上一共学习了常用的四种连接，内连接、左连接、右链接和全连接，还有两种不常用的筛选连接，分别如下：

```
## [1] "hw1_a inner join hw1_b, by ID, row count:  178"

## [1] "hw1_a left join hw1_b, by ID, row count:  189"

## [1] "hw1_a right join hw1_b, by ID, row count:  189"

## [1] "hw1_a full join hw1_b, by ID, row count:  200"

## [1] "hw1_a semi join hw1_b, by ID, row count:  178"

## [1] "hw1_a anti join hw1_b, by ID, row count:  11"
```

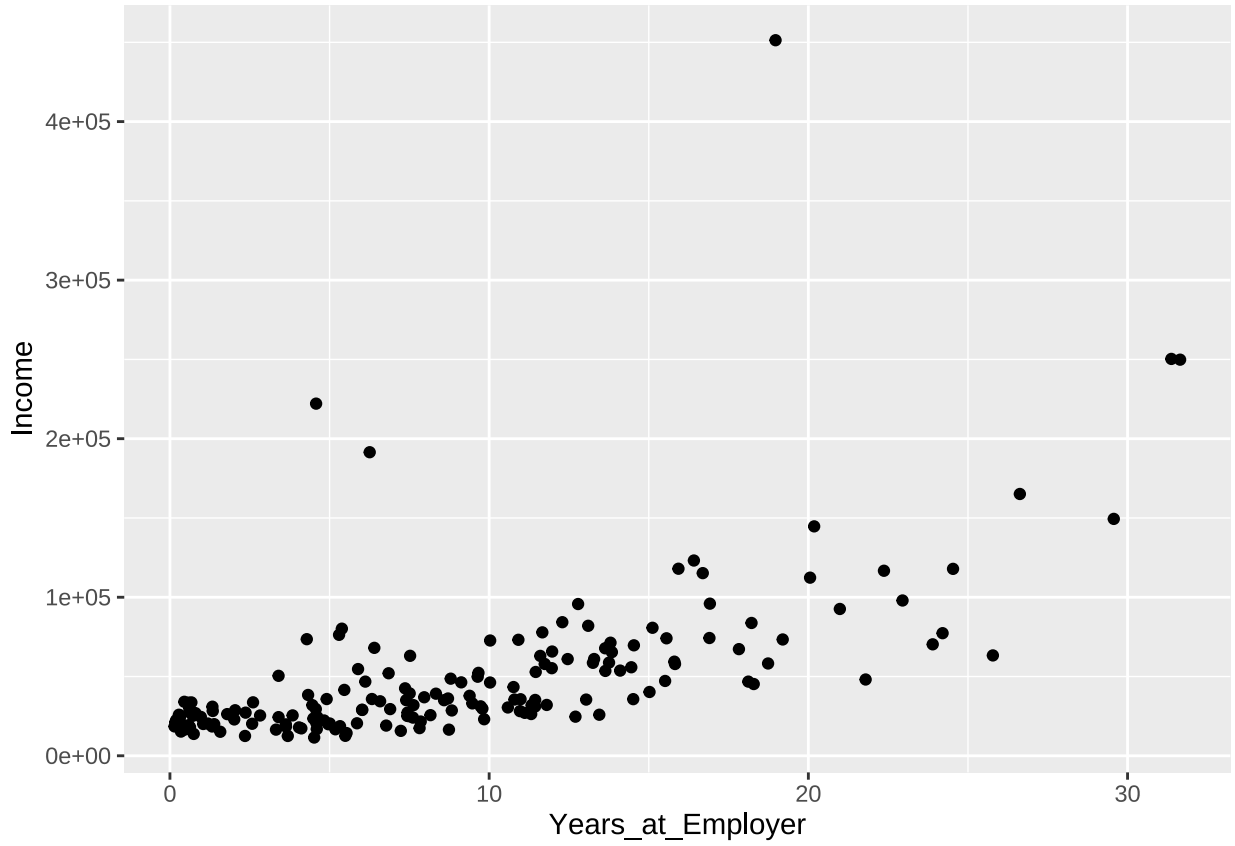
3. 请筛选出 hw1_a 中收入大于 4000 的样本，并将此样本和 hw1_b 中 Is_Default=1 的样本合并，你可以使用 inner join 的方式。这一问中你可以用 pipe 的形式

下表为结果示例

```
## # A tibble: 41 x 8
##       ID   Age Years_at_Employer Years_at_Address  Income Credit_Card_Debt
##    <dbl> <dbl>             <dbl>            <dbl>   <dbl>           <dbl>
##  1     2  34.6              12.0             1.49   65765.          -15598.
##  2     3  37.7              12.5             0.0854 61002.          -11402.
```

```
##  3     6  39.3            4.58           2.03   222106.         -16353.
##  4    11  35.3            1.04           0.776   20060.          -3899.
##  5    13  32.3            7.40           2.90    35108.          -1316.
##  6    25  49.4            4.57           0.669   29489.          -1202.
##  7    31  39.4            2.35           1.15    12508.          -3783.
##  8    39  26.5            0.746          1.53    13790.          -5586.
##  9    47  32.2            7.37           1.26    42545.          -5967.
## 10    48  29.3            4.33           1.14    38367.          -2460.
## # i 31 more rows
## # i 2 more variables: Automobile_Debt <dbl>, Is_Default <dbl>
```

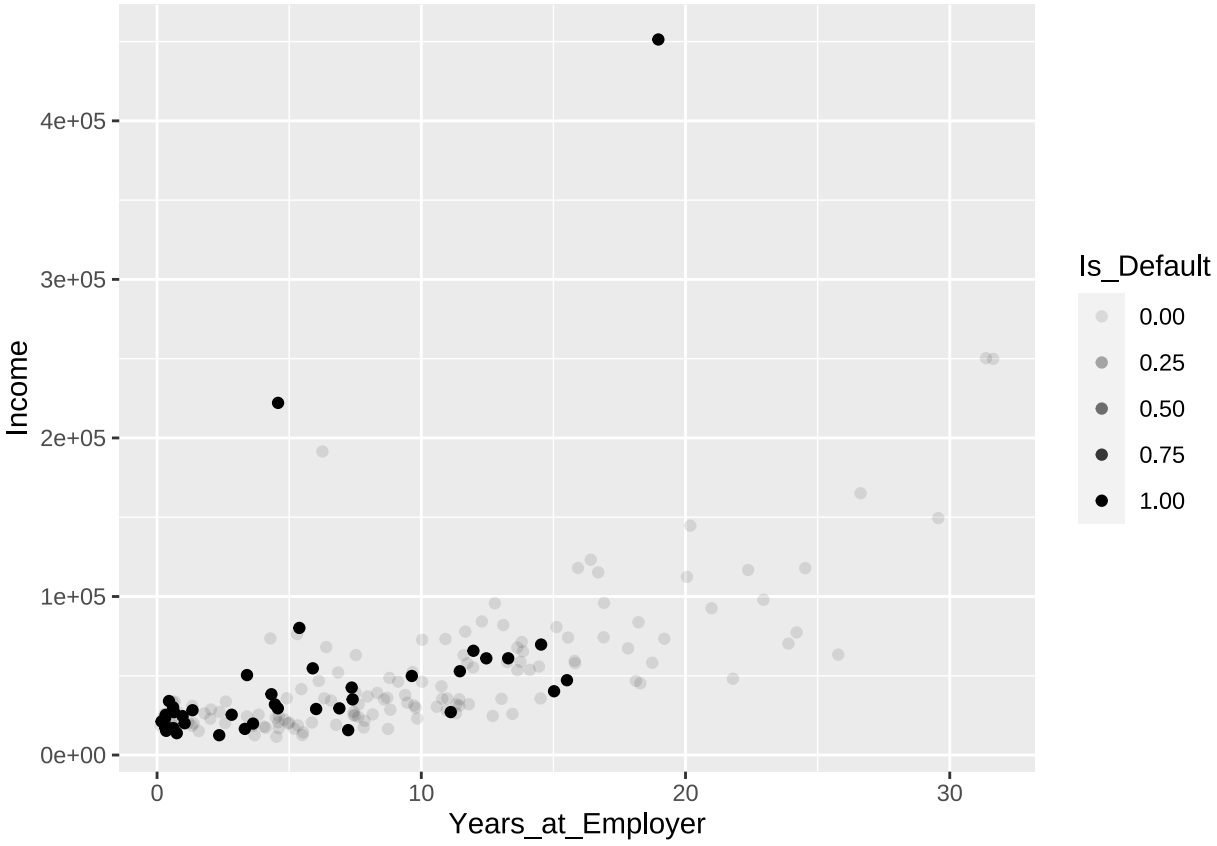
4. 在第2问的基础上, 请给出Income对Years_at_Employer的散点图，你发现了哪些趋势和现象?



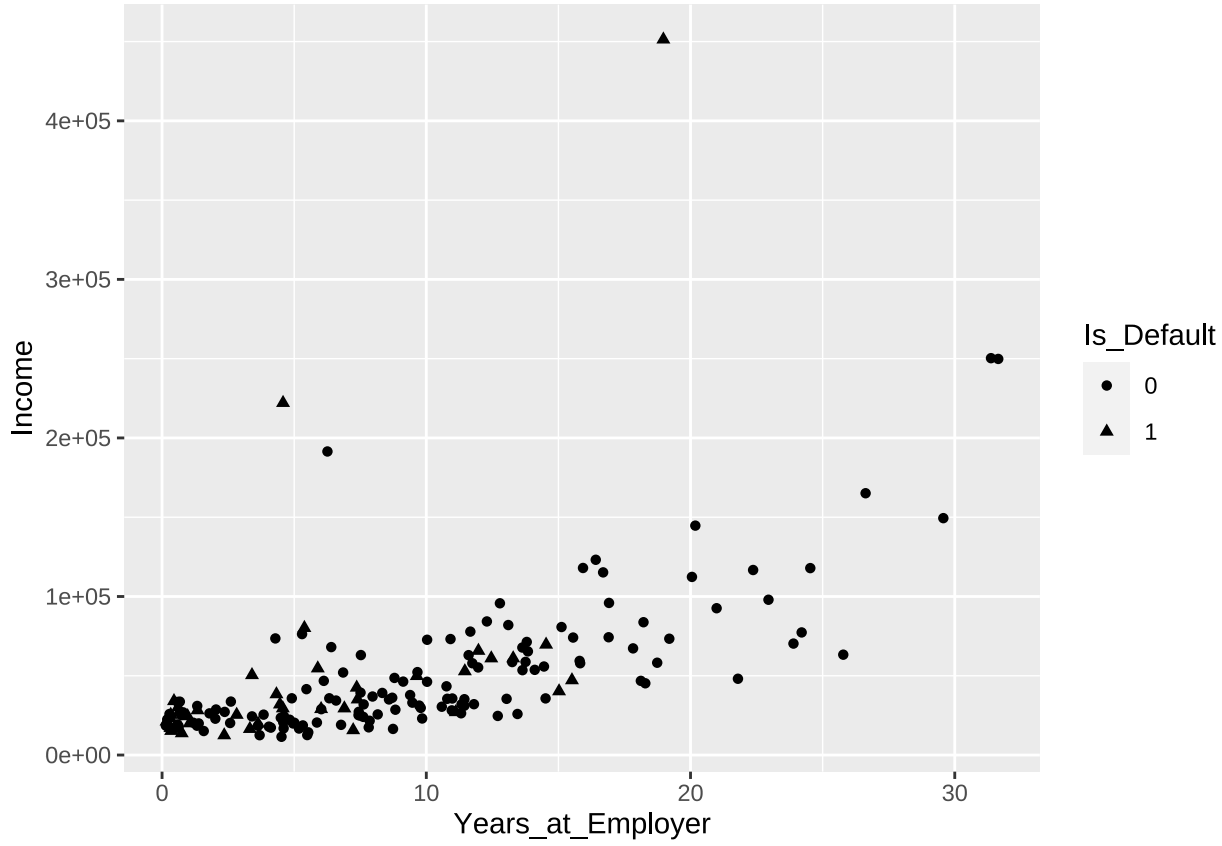
发现

- 整体上随着工作年限的增加收入程递增现象
- 10年以内增长趋势比较平缓，10年至20年增加相比于前者更加明显

5. 在第 4 问的基础上 按照 Is_Default 增加一个维度，请展示两变量在不同违约状态的散点图。请使用明暗程度作为区分方



6. 对于第5问，请使用形状(shape)作为另外一种区分方式



7. 请找出各个列的缺失值，并删除相应的行。请报告每一变量的缺失值个数，以及所有缺失值总数

如下表所示，在将hw1_a与hw1_b进行全连接后产生的缺失值个数（因为分别看每个表并没有缺失值）

```
##                ID               Age Years_at_Employer  Years_at_Address
##                 0                11                11                11
##            Income  Credit_Card_Debt   Automobile_Debt        Is_Default
##                11                11                11                11
```

```
## [1] "NA count:  77"
```

```
## [1] "remove NA, count:  178"
```

8. 找出 Income 中的极端值并滤掉对应行的数据

```
## [1] "Lower Outliers:  -30274.8140854054"
```

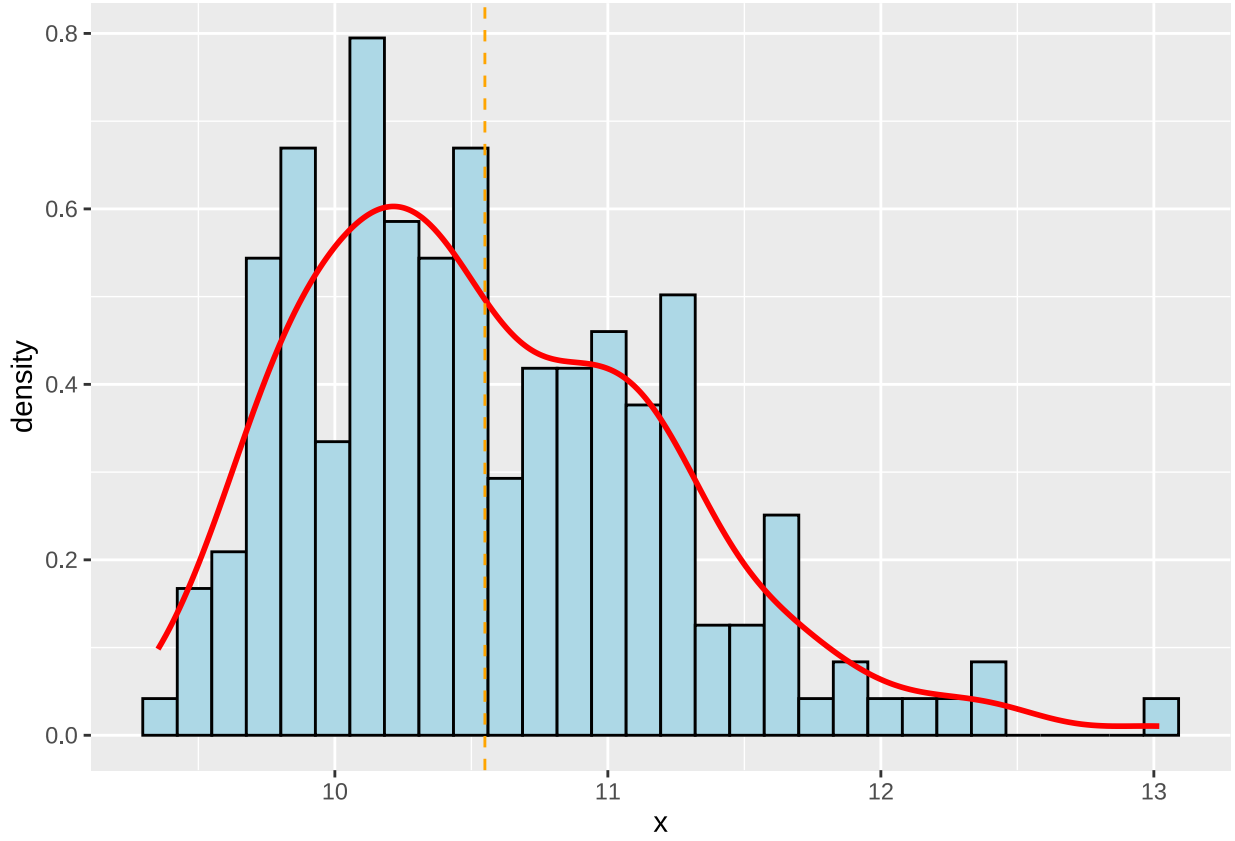
```
## [1] "High Outliers:  113144.658554217"
```

去掉极端值自后的列表还有176行，数据示例如下表：

```
## # A tibble: 176 x 5
##       ID   Age Years_at_Employer Years_at_Address Income
```

```
##    <dbl> <dbl>             <dbl>            <dbl>  <dbl>
##  1     1  32.5              9.39           0.298  37844.
##  2     2  34.6             12.0            1.49   65765.
##  3     3  37.7             12.5            0.0854 61002.
##  4     4  28.7              1.39           1.84   19953.
##  5     5  32.6              7.49           0.234  24970.
##  6     7  46.8             16.9            0.998  74283.
##  7     9  46.8             12.0            0.669  55248.
##  8    10  27.3              9.47           0.479  33040.
##  9    11  35.3              1.04           0.776  20060.
## 10    13  32.3              7.40           2.90   35108.
## # i 166 more rows
```

9. 将 Income 对数化，并画出直方图和 density curve，你有什么发现?



发现

- 整体趋势大致符合正态分布，但又不完全符合，应该是数据来源并不是完全随机抽样