第一次作业

李游游

2023-10-13

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

```
## # 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>
```

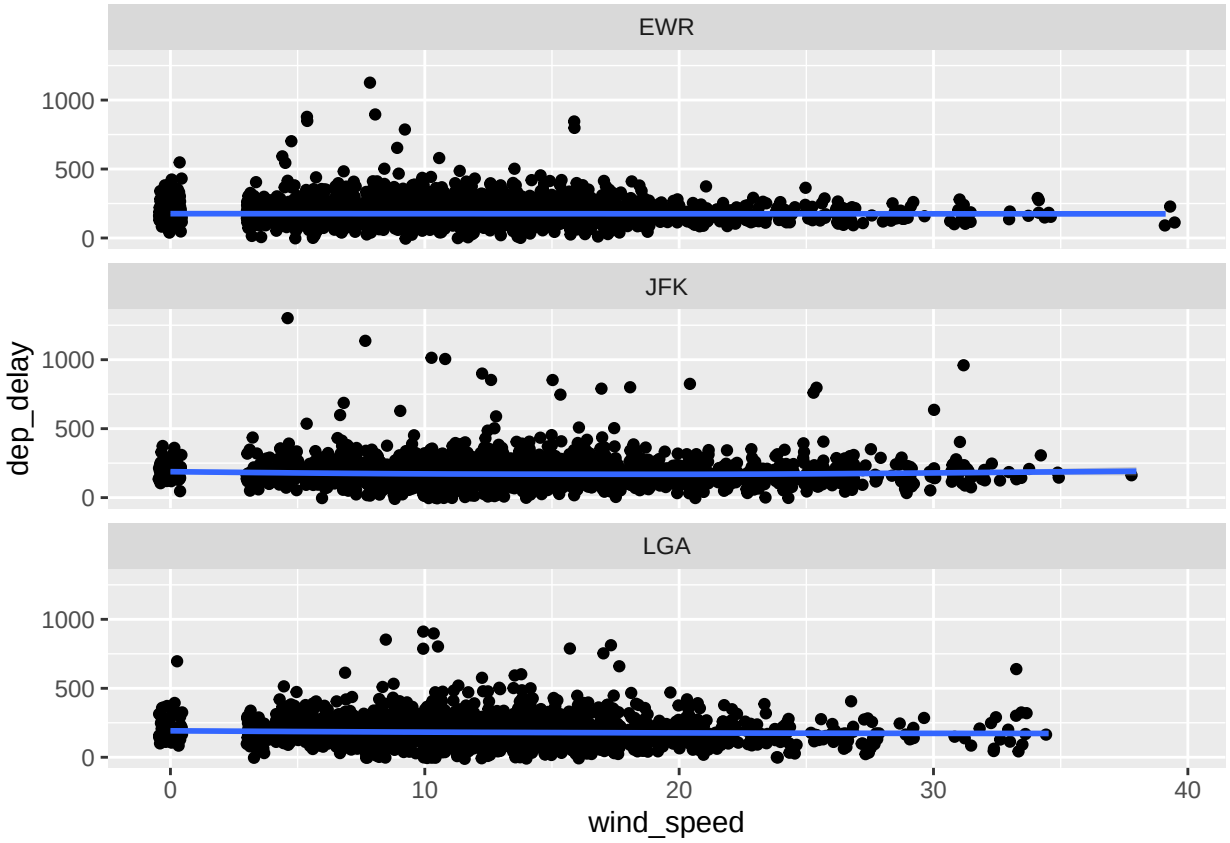
2. 将生成的 flight_arr2hr 数据集根据目的地 (dest) 进行分组，统计出抵达每个目的地的航班数量，筛选出抵达航班数量前十名的目的地，将结果命名为 top10_dest

```
## # 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

```
## # 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(y 轴) 的散点图，以及平滑曲线



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 如下表，2013 年共有 8225 个班次被取消

```
## # A tibble: 8,255 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       NA           1630        NA       NA           1815
##  2  2013     1     1       NA           1935        NA       NA           2240
##  3  2013     1     1       NA           1500        NA       NA           1825
##  4  2013     1     1       NA            600        NA       NA            901
##  5  2013     1     2       NA           1540        NA       NA           1747
##  6  2013     1     2       NA           1620        NA       NA           1746
##  7  2013     1     2       NA           1355        NA       NA           1459
##  8  2013     1     2       NA           1420        NA       NA           1644
##  9  2013     1     2       NA           1321        NA       NA           1536
## 10  2013     1     2       NA           1545        NA       NA           1910
## # i 8,245 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>
```

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量

```
## # A tibble: 16 x 3
##    carrier dest_with_most_flights flights_count_most
##    <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：数据链接及画图

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

数据 hw1_a:

变量 ID：形式为 numeric, 最小值为 1, 最大值为 200, 中值为 98, 均值为 98.8148148, 标准差为 57.3208828

变量 Age：形式为 numeric, 最小值为 20.1895762, 最大值为 55.7240627, 中值为 33.2566348, 均值为 34.9624273, 标准差为 8.254665

变量 Years_at_Employer：形式为 numeric, 最小值为 0.1434773, 最大值为 31.6460288, 中值为 7.629263, 均值为 8.9027992, 标准差为 6.835451

变量 Years_at_Address：形式为 numeric, 最小值为 0.0051076, 最大值为 3.6961479, 中值为 0.6206971, 均值为 0.784322, 标准差为 0.6362658

变量 Income：形式为 numeric, 最小值为 1.1522101×10^4, 最大值为 4.5131967×10^5, 中值为 3.4375085×10^4, 均值为 4.9626064×10^4, 标准差为 4.9034311×10^4

数据 hw1_b:

变量 ID：形式为 numeric, 最小值为 1, 最大值为 200, 中值为 100, 均值为 101.4603175, 标准差为 57.9657342

变量 Credit_Card_Debt：形式为 numeric, 最小值为 -3.2050377×10^4, 最大值为 34.1638172, 中值为-1833.3318932, 均值为-3287.1107542, 标准差为 3972.9489713

变量 Automobile_Debit：形式为 NULL, 最小值为 ∞, 最大值为 $-\infty$, 中值为, 均值为 NA, 标准差为 NA

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

```
## 内连接 inner join，通过ID变量连接，总行数： 189
```

```
## 左连接 left join，通过ID变量连接，总行数： 189
```

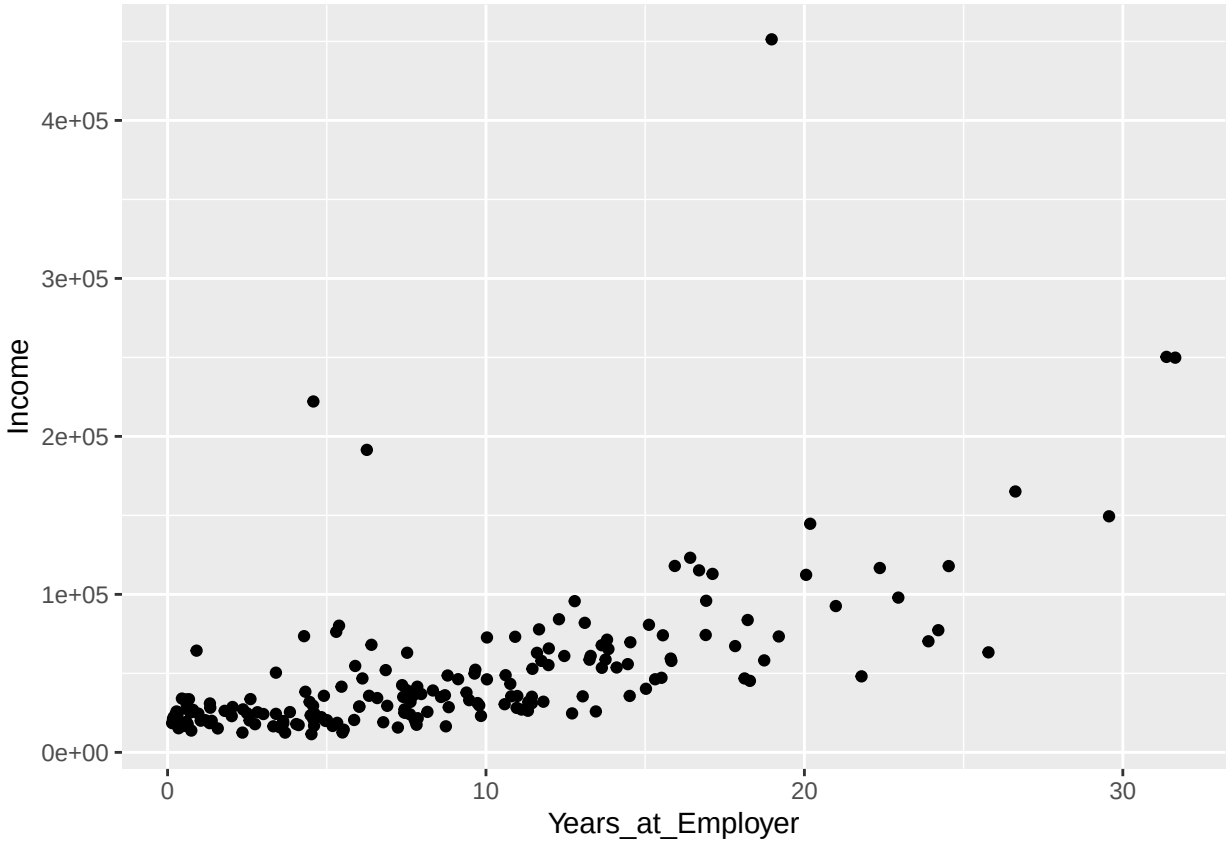
```
## 右连接 right join，通过ID变量连接，总行数： 189
```

```
## 全连接 full join，通过ID变量连接，总行数： 189
```

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

```
## # 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>
```

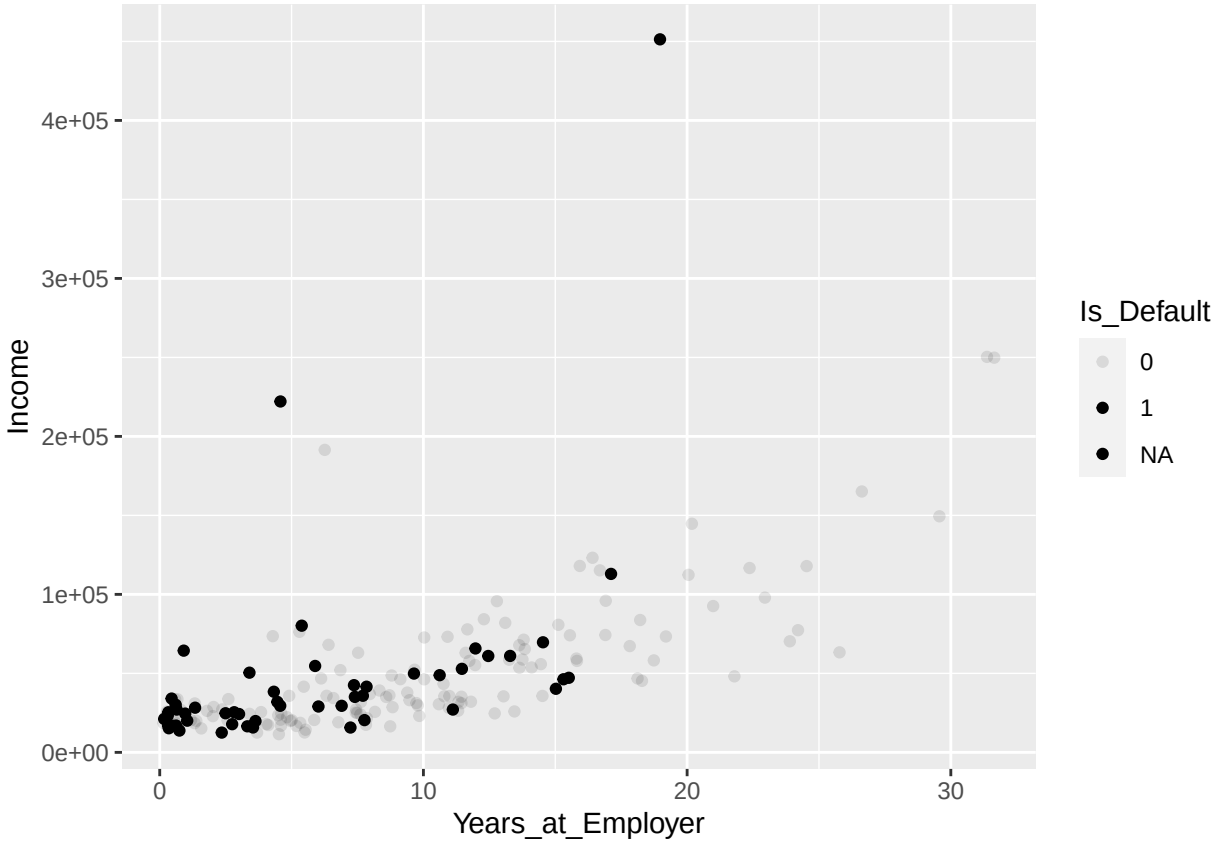
4. 在第 2 问的基础上, 请给出 Income 对 Years_at_Employer 的散点图，你发现了哪些趋势和现象?



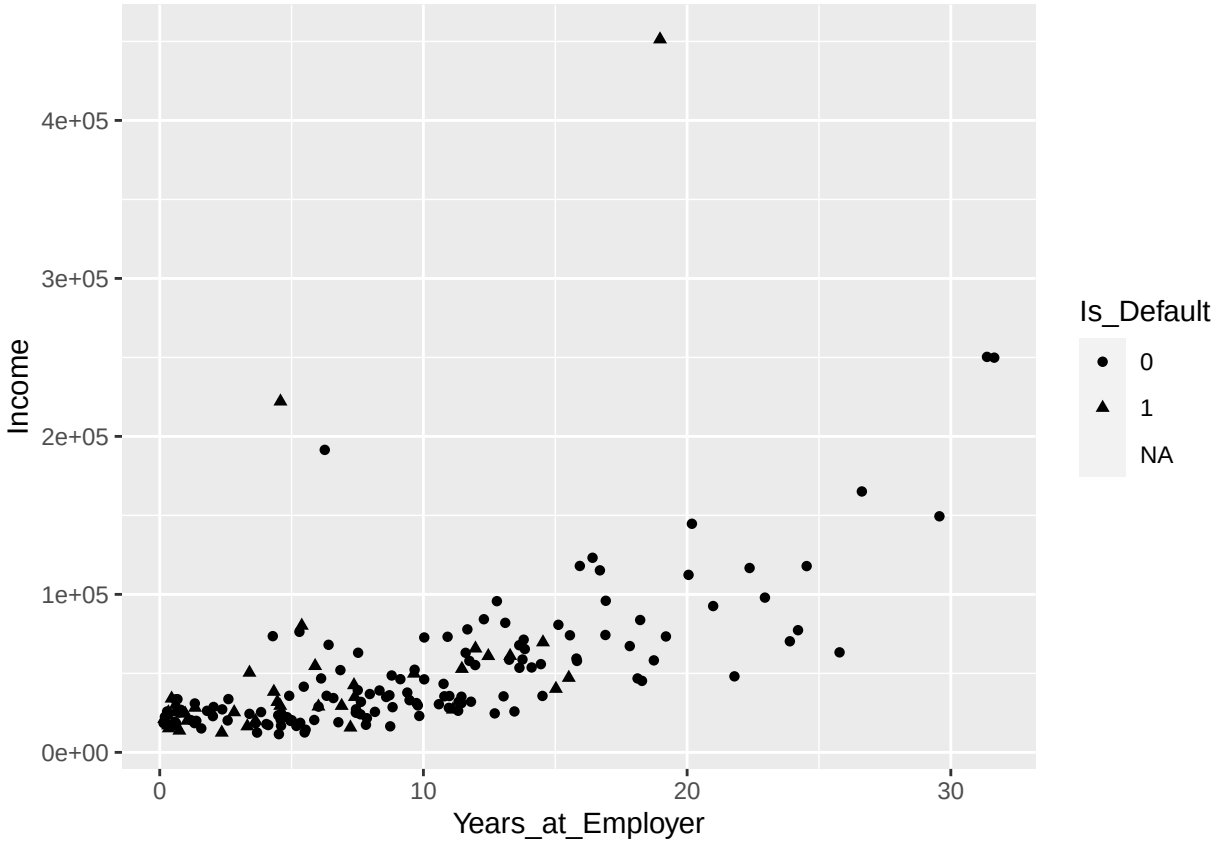
发现

- 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式



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

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

```
## [1] "所有的缺失值数量: 33"
```

```
## [1] "删除各行缺失值后的记录数:  178"
```

8. 找出 Income 中的极端值并滤掉对应行的数据

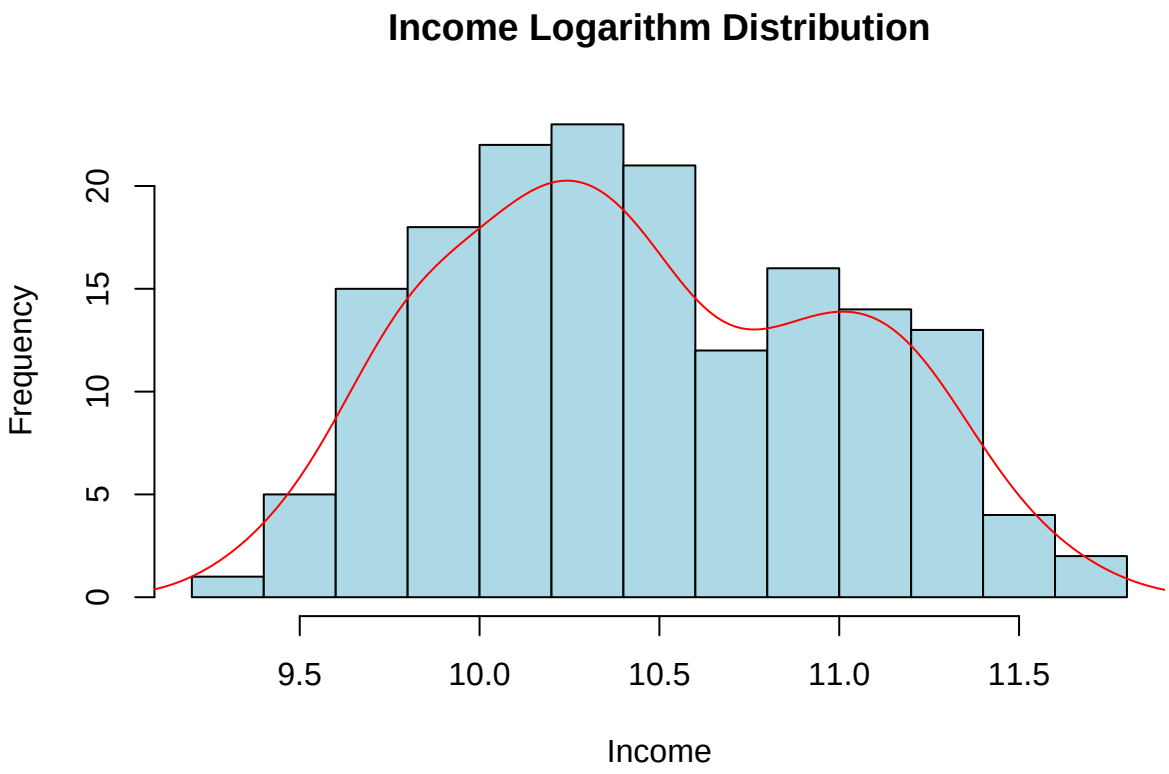
```
## 低处利群点:  -31760.17     当前数据集中不存。
```

```
## 高处离群点:  116003.8
```

```
## # A tibble: 166 x 8
##       ID   Age Years_at_Employer Years_at_Address Income Credit_Card_Debt
##    <dbl> <dbl>             <dbl>            <dbl>  <dbl>            <dbl>
##  1     1  32.5              9.39           0.298  37844.           -3247.
##  2     2  34.6             12.0            1.49   65765.          -15598.
##  3     3  37.7             12.5            0.0854 61002.          -11402.
##  4     4  28.7              1.39           1.84   19953.           -1233.
```

```
##  5     5  32.6              7.49          0.234  24970.          -1136.
##  6     7  46.8             16.9           0.998  74283.          -4468.
##  7     9  46.8             12.0           0.669  55248.          -7435.
##  8    10  27.3              9.47          0.479  33040.          -1833.
##  9    11  35.3              1.04          0.776  20060.          -3899.
## 10    13  32.3              7.40          2.90   35108.          -1316.
## # i 156 more rows
## # i 2 more variables: Automobile_Debt <dbl>, Is_Default <dbl>
```

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