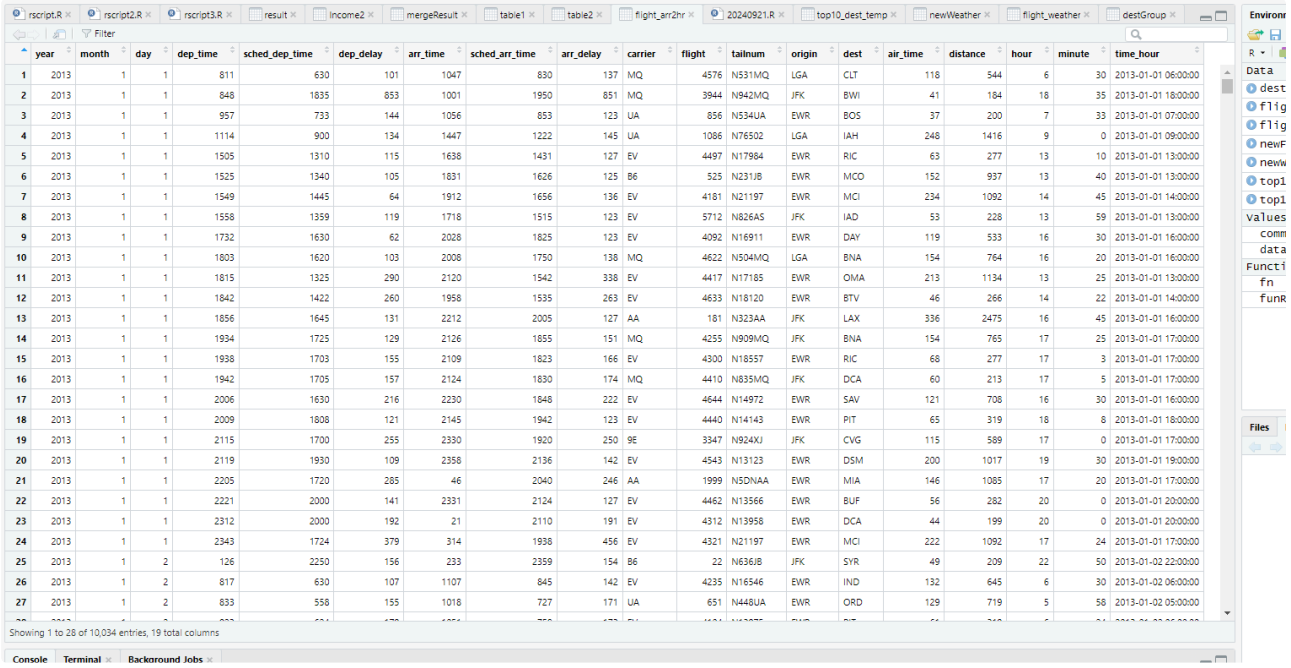
R 语言作业

第一大题

1. 第一问结果：

```R
flight_arr2hr = flights %>% filter(arr_delay > 120)
```



	year	month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carrier	flight	tailnum	origin	dest	air_time	distance	hour	minute	time_hour
1	2013	1	1	811	630	101	1047	830	137	MQ	4576	N531MQ	LGA	CLT	118	544	6	30	2013-01-01 06:00:00
2	2013	1	1	848	1835	853	1001	1950	851	MQ	3944	N942MQ	JFK	BWI	41	184	18	35	2013-01-01 18:00:00
3	2013	1	1	957	733	144	1056	853	123	UA	856	N534UA	EWR	BOS	37	200	7	33	2013-01-01 07:00:00
4	2013	1	1	1114	900	134	1447	1222	145	UA	1086	N76502	LGA	IAH	248	1416	9	0	2013-01-01 09:00:00
5	2013	1	1	1505	1310	115	1638	1431	127	EV	4497	N17984	EWR	RIC	63	277	13	10	2013-01-01 13:00:00
6	2013	1	1	1525	1340	105	1831	1626	125	B6	525	N231JB	EWR	MCO	152	937	13	40	2013-01-01 13:00:00
7	2013	1	1	1549	1445	64	1912	1656	136	EV	4181	N21197	EWR	MCI	234	1092	14	45	2013-01-01 14:00:00
8	2013	1	1	1558	1359	119	1718	1515	123	EV	5712	N826AS	JFK	IAD	53	228	13	59	2013-01-01 13:00:00
9	2013	1	1	1732	1630	62	2028	1825	123	EV	4092	N16911	EWR	DAY	119	533	16	30	2013-01-01 16:00:00
10	2013	1	1	1803	1620	103	2006	1750	136	MQ	4622	N504MQ	LGA	BNA	154	764	16	20	2013-01-01 16:00:00
11	2013	1	1	1815	1325	290	2120	1542	338	EV	4417	N17185	EWR	OMA	213	1134	13	25	2013-01-01 13:00:00
12	2013	1	1	1842	1422	260	1958	1535	263	EV	4633	N18120	EWR	BTV	46	266	14	22	2013-01-01 14:00:00
13	2013	1	1	1856	1645	131	2212	2005	127	AA	181	N323AA	JFK	LAX	336	2475	16	45	2013-01-01 16:00:00
14	2013	1	1	1934	1725	129	2126	1855	151	MQ	4255	N909MQ	JFK	BNA	154	765	17	25	2013-01-01 17:00:00
15	2013	1	1	1938	1703	155	2109	1823	166	EV	4300	N18557	EWR	RIC	68	277	17	3	2013-01-01 17:00:00
16	2013	1	1	1942	1705	157	2124	1830	174	MQ	4410	N835MQ	JFK	DCA	60	213	17	5	2013-01-01 17:00:00
17	2013	1	1	2006	1630	216	2230	1848	222	EV	4644	N14972	EWR	SAV	121	708	16	30	2013-01-01 16:00:00
18	2013	1	1	2009	1808	121	2145	1942	123	EV	4440	N14143	EWR	PIT	65	319	18	8	2013-01-01 18:00:00
19	2013	1	1	2115	1700	255	2330	1920	250	9E	3347	N924XJ	JFK	CVG	115	589	17	0	2013-01-01 17:00:00
20	2013	1	1	2119	1930	109	2358	2136	142	EV	4543	N13123	EWR	DSM	200	1017	19	30	2013-01-01 19:00:00
21	2013	1	1	2205	1720	285	46	2040	246	AA	1999	N5DNAA	EWR	MIA	146	1085	17	20	2013-01-01 17:00:00
22	2013	1	1	2221	2000	141	2331	2124	127	EV	4462	N13566	EWR	BUF	56	282	20	0	2013-01-01 20:00:00
23	2013	1	1	2312	2000	192	21	2110	191	EV	4312	N13958	EWR	DCA	44	199	20	0	2013-01-01 20:00:00
24	2013	1	1	2343	1724	379	314	1938	456	EV	4321	N21197	EWR	MCI	222	1092	17	24	2013-01-01 17:00:00
25	2013	1	2	126	2250	156	233	2359	154	B6	22	N636JB	JFK	SYR	49	209	22	50	2013-01-02 22:00:00
26	2013	1	2	817	630	107	1107	845	142	EV	4235	N16546	EWR	IND	132	645	6	30	2013-01-02 06:00:00
27	2013	1	2	833	558	155	1018	727	171	UA	651	N448UA	EWR	ORD	129	719	5	58	2013-01-02 05:00:00

Showing 1 to 28 of 10,034 entries, 19 total columns

2. 第二问结果：

```R
top10_dest  = flight_arr2hr %>% group_by(dest) %>% summarise(count=n()) %>% arrange(desc(count)) %>% head(10)
```

	dest	count
1	ATL	572
2	ORD	570
3	SFO	405
4	MCO	384
5	FLL	375
6	CLT	361
7	BOS	349
8	LAX	312
9	DTW	270
10	IAD	265

3. 第三问结果：

R

```r
newWeather = weather %>% select(origin:hour, humid, wind_speed)

commonCol = intersect(names(newWeather), names(flight_arr2hr))

flight_weather = left_join(newWeather, flight_arr2hr, by = commonCol)
```

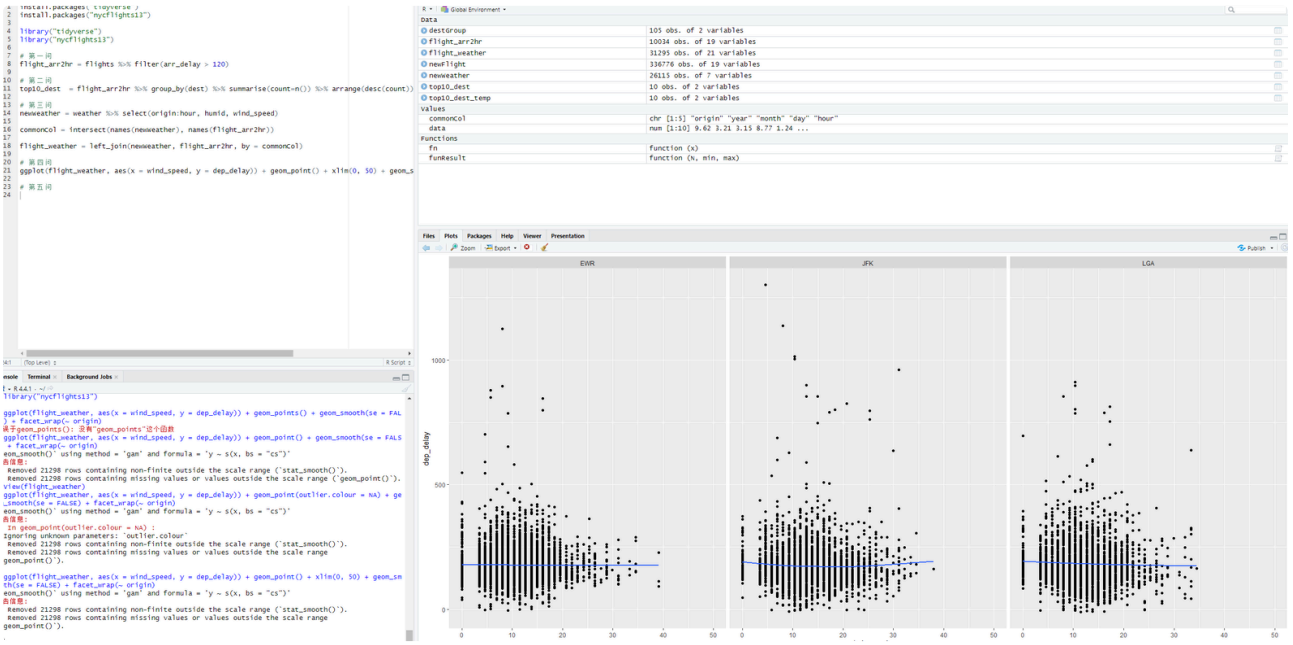
Filter

	origin	year	month	day	hour	humid	wind_speed	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carrier	flight	tailnum	dest	air_time	distance	minute
1	EWR	2013	1	1	1	59.37	10.35702	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
2	EWR	2013	1	1	2	61.63	8.05546	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
3	EWR	2013	1	1	3	64.43	11.50780	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
4	EWR	2013	1	1	4	62.21	12.65858	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
5	EWR	2013	1	1	5	64.43	12.65858	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
6	EWR	2013	1	1	6	67.21	11.50780	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
7	EWR	2013	1	1	7	64.43	14.96014	957	733	144	1056	853	123	UA	856	N534UA	BOS	37	200	
8	EWR	2013	1	1	8	62.21	10.35702	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
9	EWR	2013	1	1	9	62.21	14.96014	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
10	EWR	2013	1	1	10	59.65	13.80936	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
11	EWR	2013	1	1	11	57.06	14.96014	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
12	EWR	2013	1	1	13	69.67	16.11092	1505	1310	115	1638	1431	127	EV	4497	N17984	RIC	63	277	
13	EWR	2013	1	1	13	69.67	16.11092	1525	1340	105	1831	1626	125	B6	525	N231JB	MCO	152	937	
14	EWR	2013	1	1	13	69.67	16.11092	1815	1325	290	2120	1542	338	EV	4417	N17185	OMA	213	1134	
15	EWR	2013	1	1	14	54.68	13.80936	1549	1445	64	1912	1656	136	EV	4181	N21197	MCI	234	1092	
16	EWR	2013	1	1	14	54.68	13.80936	1842	1422	260	1958	1535	263	EV	4633	N18120	BTV	46	266	
17	EWR	2013	1	1	15	57.04	9.20624	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
18	EWR	2013	1	1	16	49.62	13.80936	1732	1630	62	2028	1825	123	EV	4092	N16911	DAY	119	533	
19	EWR	2013	1	1	16	49.62	13.80936	2006	1630	216	2230	1848	222	EV	4644	N14972	SAV	121	708	
20	EWR	2013	1	1	17	49.83	11.50780	1938	1703	155	2109	1823	166	EV	4300	N18557	RIC	68	277	
21	EWR	2013	1	1	17	49.83	11.50780	2205	1720	285	46	2040	246	AA	1999	N5DNAA	MIA	146	1085	
22	EWR	2013	1	1	17	49.83	11.50780	2343	1724	379	314	1938	456	EV	4321	N21197	MCI	222	1092	
23	EWR	2013	1	1	18	45.43	12.65858	2009	1808	121	2145	1942	123	EV	4440	N14143	PIT	65	319	
24	EWR	2013	1	1	19	42.84	10.35702	2119	1930	109	2358	2136	142	EV	4543	N13123	DSM	200	1017	
25	EWR	2013	1	1	20	49.19	14.96014	2221	2000	141	2331	2124	127	EV	4462	N13566	BUF	56	282	
26	EWR	2013	1	1	20	49.19	14.96014	2312	2000	192	21	2110	191	EV	4312	N13958	DCA	44	199	
27	EWR	2013	1	1	21	48.48	18.41248	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	

4. 第四问结果：

SQL

```
ggplot(flight_weather, aes(x = wind_speed, y = dep_delay)) + geom_point() + xlim(0, 50)
+ geom_smooth(se = FALSE) + facet_wrap(~ origin)
```



5. 第五问结果:

R

```
flights %>% filter(is.na(dep_time)) %>% group_by(carrier) %>% summarise(count = n())
```

```
ggplot(flight_weather, aes(x = wind_speed, y = dep_delay)) + geom_point() + xlim(0, 50) + geom_s

# 第五问
flights %>% filter(is.na(dep_time)) %>% group_by(carrier) %>% summarise(count = n())
```

24:1 (Top Level) R Script

Console Terminal Background Jobs

R 4.4.1 · ~/

```
2: Removed 21298 rows containing missing values or values outside the scale range
(`geom_point()`).
>
> Length(flights %>% filter(is.na(dep_time)))
错误于Length(flights %>% filter(is.na(dep_time))):
  没有"Length"这个函数
> length(flights %>% filter(is.na(dep_time)))
[1] 19
> View(flights %>% filter(is.na(dep_time)))
> View(flights %>% filter(is.na(dep_time)))
> View(flights %>% filter(is.na(dep_time)))
> flights %>% filter(is.na(dep_time)) %>% group_by(carrier) %>% summarise(count = n())
# A tibble: 15 × 2
   carrier 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. 第六问结果:

SQL

```
result = flights %>% group_by(carrier, dest) %>% summarise(count = n())
```

	carrier	dest	count
1	9E	ATL	59
2	9E	AUS	2
3	9E	AVL	10
4	9E	BGR	1
5	9E	BNA	474
6	9E	BOS	914
7	9E	BTV	2
8	9E	BUF	833
9	9E	BWI	856
10	9E	CAE	3
11	9E	CHS	348
12	9E	CLE	349
13	9E	CLT	291
14	9E	CMH	13
15	9E	CVG	1559
16	9E	DAY	391
17	9E	DCA	1074
18	9E	DFW	379
19	9E	DSM	91
20	9E	DTW	1013
21	9E	GRR	44
22	9E	GSO	1
23	9E	GSP	102
24	9E	IAD	664
25	9E	IND	401
26	9E	JAX	400
27	9E	LEX	1

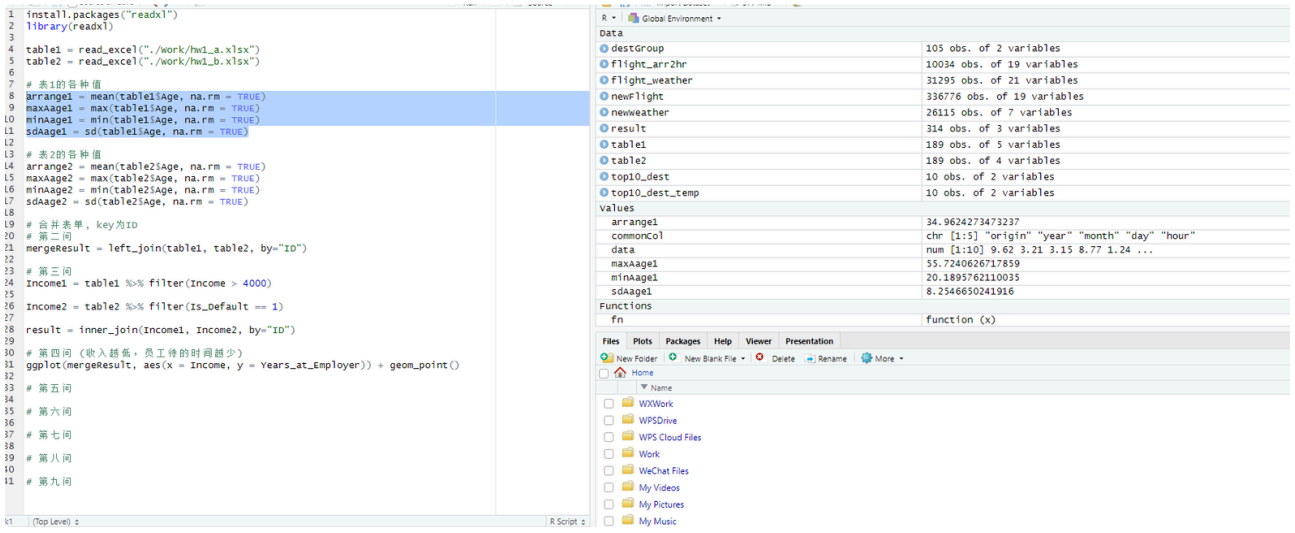
Showing 1 to 28 of 314 entries, 3 total columns

第二大题

1. 第一问

```R
install.packages("readxl")
library(readxl)

table1 = read_excel("./Work/hw1_a.xlsx")
table2 = read_excel("./Work/hw1_b.xlsx")

# 表1的各种值
arrange1 = mean(table1$Age, na.rm = TRUE)
maxAage1 = max(table1$Age, na.rm = TRUE)
minAage1 = min(table1$Age, na.rm = TRUE)
sdAage1 = sd(table1$Age, na.rm = TRUE)

# 表2的各种值
arrange2 = mean(table2$Age, na.rm = TRUE)
```

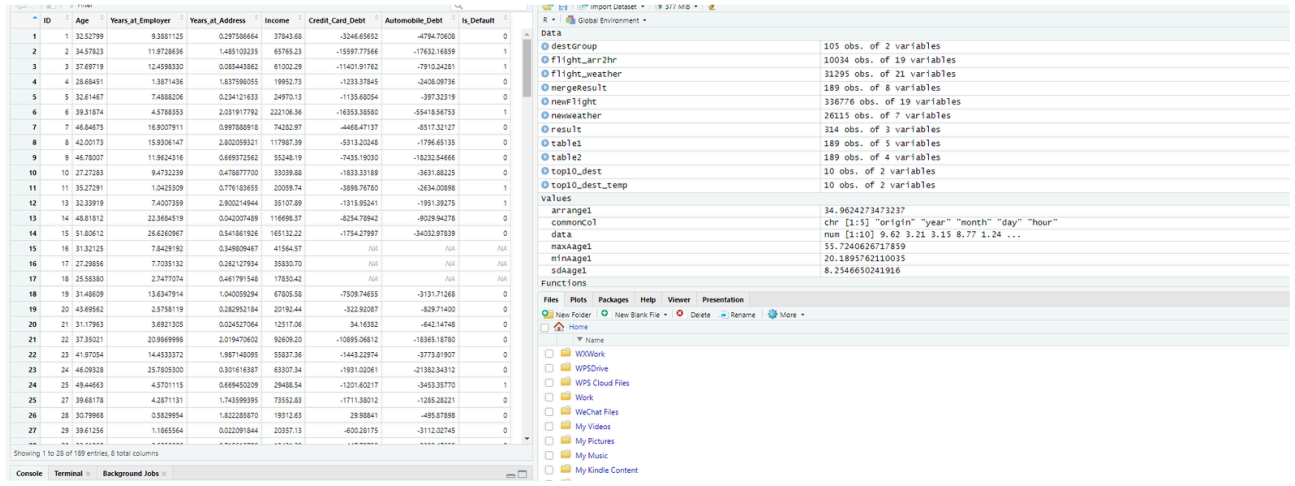
```r
maxAage2 = max(table2$Age, na.rm = TRUE)
minAage2 = min(table2$Age, na.rm = TRUE)
sdAage2 = sd(table2$Age, na.rm = TRUE)
```



2. 第二问

R

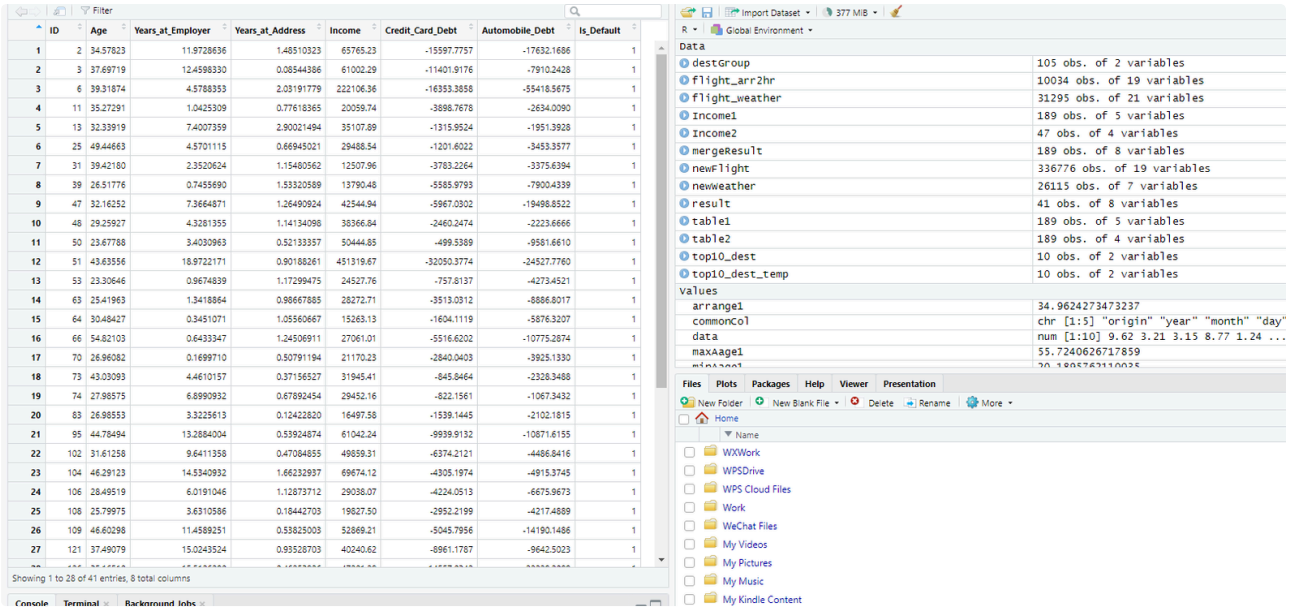
```r
mergeResult = left_join(table1, table2, by="ID")
```



3. 第三问

R

```r
Income1 = table1 %>% filter(Income > 4000)

Income2 = table2 %>% filter(Is_Default == 1)

result = inner_join(Income1, Income2, by="ID")
```



ID	Age	Years_at_Employer	Years_at_Address	Income	Credit_Card_Debt	Automobile_Debt	Is_Default
2	34.57823	11.9728636	1.48510323	65765.23	-15597.7757	-17632.1686	1
3	37.69719	12.4598330	0.08544386	61002.29	-11401.9176	-7910.2428	1
6	39.31874	4.5788353	2.03191779	222106.36	-16353.3858	-55418.5675	1
11	35.27291	1.0425309	0.77618365	20059.74	-3898.7678	-2634.0090	1
13	32.33919	7.4007359	2.90021494	35107.89	-1315.9524	-1951.3928	1
25	49.44663	4.5701115	0.66945021	29488.54	-1201.6022	-3453.3577	1
31	39.42180	2.3520624	1.15480562	12507.96	-3783.2264	-3375.6394	1
39	26.51776	0.7455690	1.53320589	13790.48	-5585.9793	-7900.4339	1
47	32.16252	7.3664871	1.26490924	42544.94	-5967.0302	-19498.8522	1
48	29.25927	4.3281355	1.14134098	38366.84	-2460.2474	-2223.6666	1
50	23.67788	3.4030963	0.52133357	50444.85	-499.5389	-9581.6610	1
51	43.63556	18.9722171	0.90188261	451319.67	-32050.3774	-24527.7760	1
53	23.30646	0.9674839	1.17299475	24527.76	-757.8137	-4273.4521	1
63	25.41963	1.3418864	0.98667685	28272.71	-3513.0312	-8886.8017	1
64	30.48427	0.3451071	1.05560667	15263.13	-1604.1119	-5876.3207	1
66	54.82103	0.6433347	1.24506911	27061.01	-5516.6202	-10775.2874	1
70	26.96082	0.1699710	0.50791194	21170.23	-2840.0403	-3925.1330	1
73	43.03093	4.4610157	0.37156527	31945.41	-845.8464	-2328.3468	1
74	27.98575	6.8990932	0.67892454	29452.16	-822.1561	-1067.3432	1
83	26.98553	3.3225613	0.12422820	16497.58	-1539.1445	-2102.1815	1
95	44.78494	13.2684004	0.53924874	61042.24	-9939.9132	-10671.6155	1
102	31.61258	9.6411358	0.47084855	49859.31	-6374.2121	-4486.8416	1
104	46.29123	14.5340932	1.66232937	69674.12	-4305.1974	-4915.3745	1
106	28.49519	6.0191046	1.12873712	29038.07	-4224.0513	-6675.9673	1
108	25.79975	3.6310586	0.18442703	19827.50	-2952.2199	-4217.4889	1
109	46.60298	11.4589251	0.53825003	52869.21	-5045.7956	-14190.1486	1
121	37.49079	15.0243524	0.93528703	40240.62	-8961.1787	-9642.5023	1

Showing 1 to 28 of 41 entries, 8 total columns

Console Terminal Background Jobs

Data

destGroup	105 obs. of 2 variables
flight_arr2hr	10034 obs. of 19 variables
flight_weather	31295 obs. of 21 variables
Income1	189 obs. of 5 variables
Income2	47 obs. of 4 variables
mergeResult	189 obs. of 8 variables
newFlight	336776 obs. of 19 variables
newWeather	26115 obs. of 7 variables
result	41 obs. of 8 variables
table1	189 obs. of 5 variables
table2	189 obs. of 4 variables
top10_dest	10 obs. of 2 variables
top10_dest_temp	10 obs. of 2 variables

Values

arrange1	34.9624273473237
commonCol	chr [1:5] "origin" "year" "month" "day"
data	num [1:10] 9.62 3.21 3.15 8.77 1.24 ...
maxAage1	55.7240626717859

Files Plots Packages Help Viewer Presentation

New Folder New Blank File Delete Rename More

Home

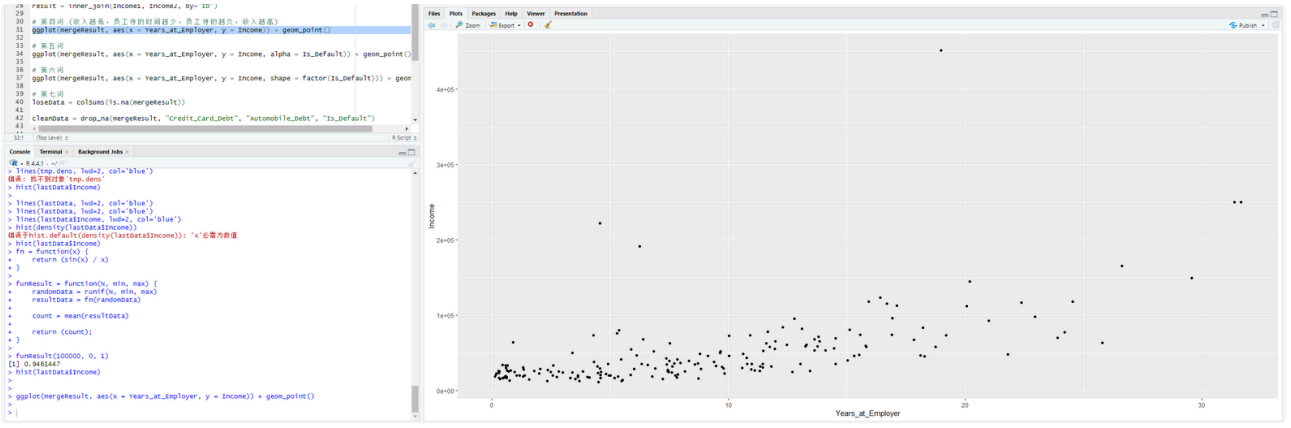
Name

WXWork
WPSDrive
WPS Cloud Files
Work
WeChat Files
My Videos
My Pictures
My Music
My Kindle Content

4. 第四问

```R
ggplot(mergeResult, aes(x = Years_at_Employer, y = Income)) + geom_point()
```

随着员工待的时间的增长，员工的收入是在增加的

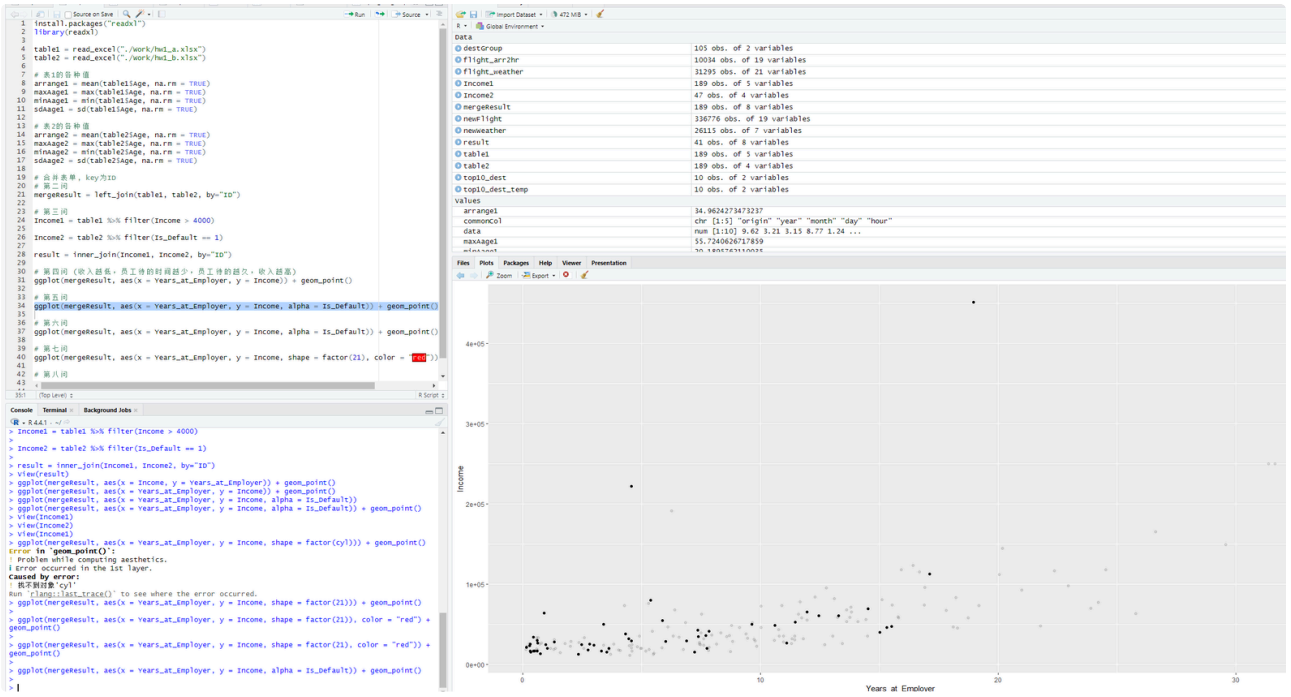


```
result = inner_join(income1, income2, by= 'ID')
# 第四问（收入越低，员工待的时间越少，员工待的越久，收入越高）
ggplot(mergeResult, aes(x = Years_at_Employer, y = Income)) + geom_point()
# 第五问
ggplot(mergeResult, aes(x = Years_at_Employer, y = Income, alpha = Is_Default)) + geom_point()
# 第六问
ggplot(mergeResult, aes(x = Years_at_Employer, y = Income, shape = factor(Is_Default))) + geom
# 第七问
loseData = colSums(is.na(mergeResult))
cleanData = drop_na(mergeResult, "Credit_Card_Debt", "Automobile_Debt", "Is_Default")
```

```
> lines(tmp.dens, lwd=2, col='blue')
错误：找不到对象'tmp.dens'
> hist(lastData$Income)
>
> lines(lastData, lwd=2, col='blue')
> lines(lastData, lwd=2, col='blue')
> lines(lastData$Income, lwd=2, col='blue')
> hist(density(lastData$Income))
错误于hist.default(density(lastData$Income)): 'x'必须为数值
> hist(lastData$Income)
> fn = function(x) {
+     return (sin(x) / x)
+ }
>
> funResult = function(n, min, max) {
+     randomData = runif(n, min, max)
+     resultData = fn(randomData)
+
+     count = mean(resultData)
+
+     return (count);
+ }
>
> funResult(100000, 0, 1)
[1] 0.946147
> hist(lastData$Income)
>
>
> ggplot(mergeResult, aes(x = Years_at_Employer, y = Income)) + geom_point()
>
```

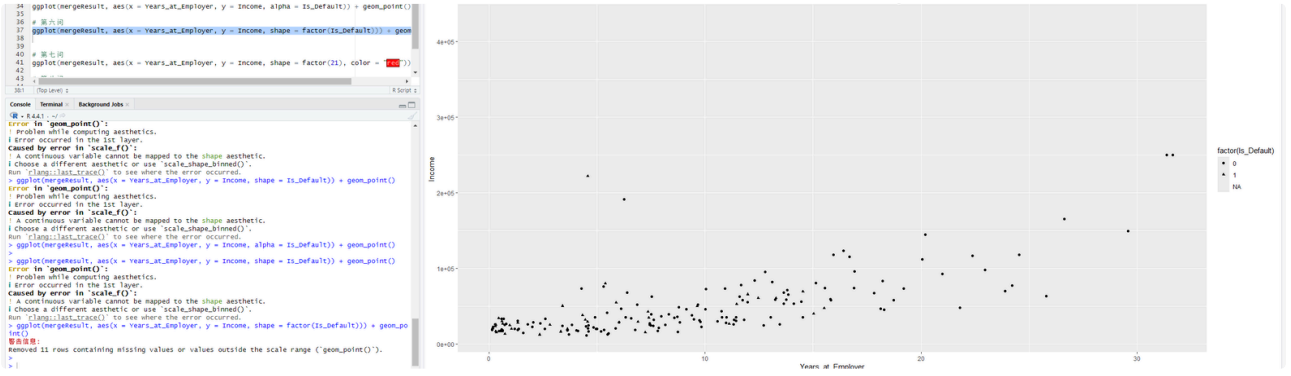

5. 第五问

```R
ggplot(mergeResult, aes(x = Years_at_Employer, y = Income, alpha = Is_Default)) + geom_
point()
```



6. 第六问

```R
ggplot(mergeResult, aes(x = Years_at_Employer, y = Income, shape = factor(Is_Default)))
+ geom_point()
```



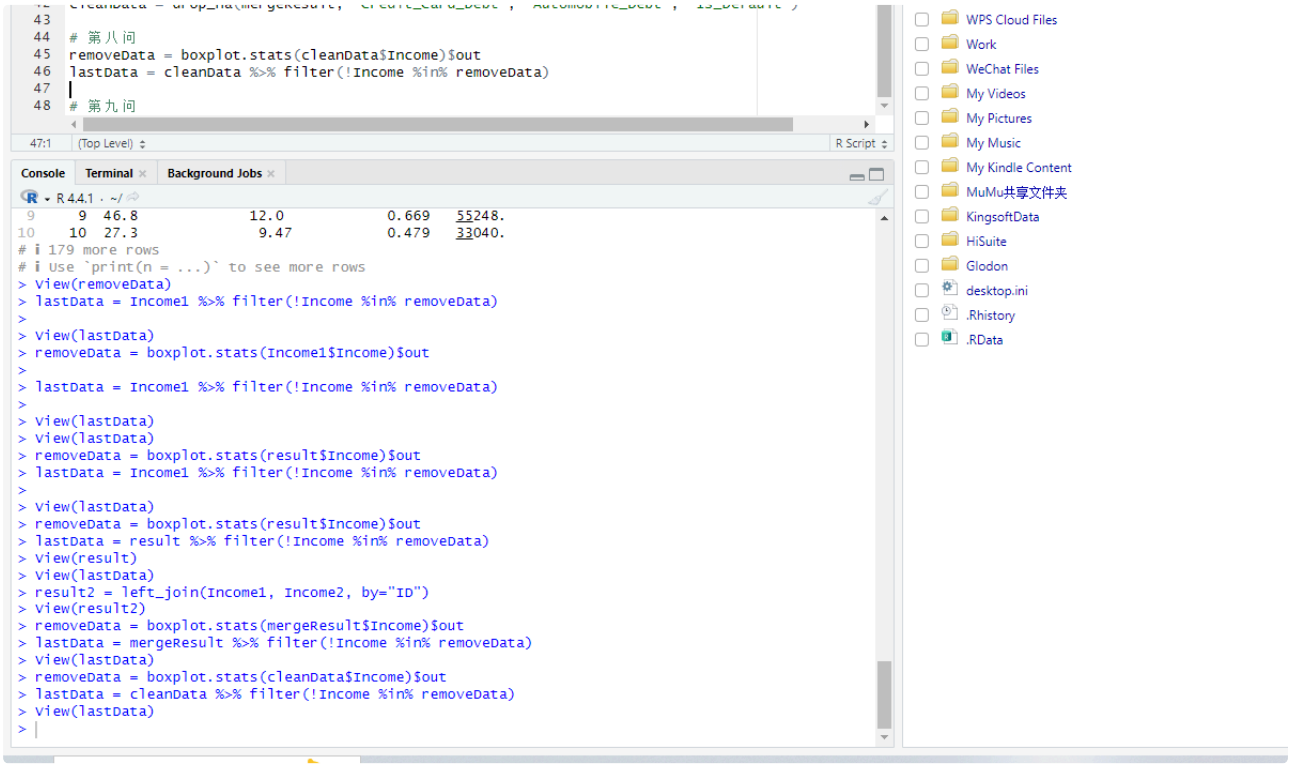
7. 第七问

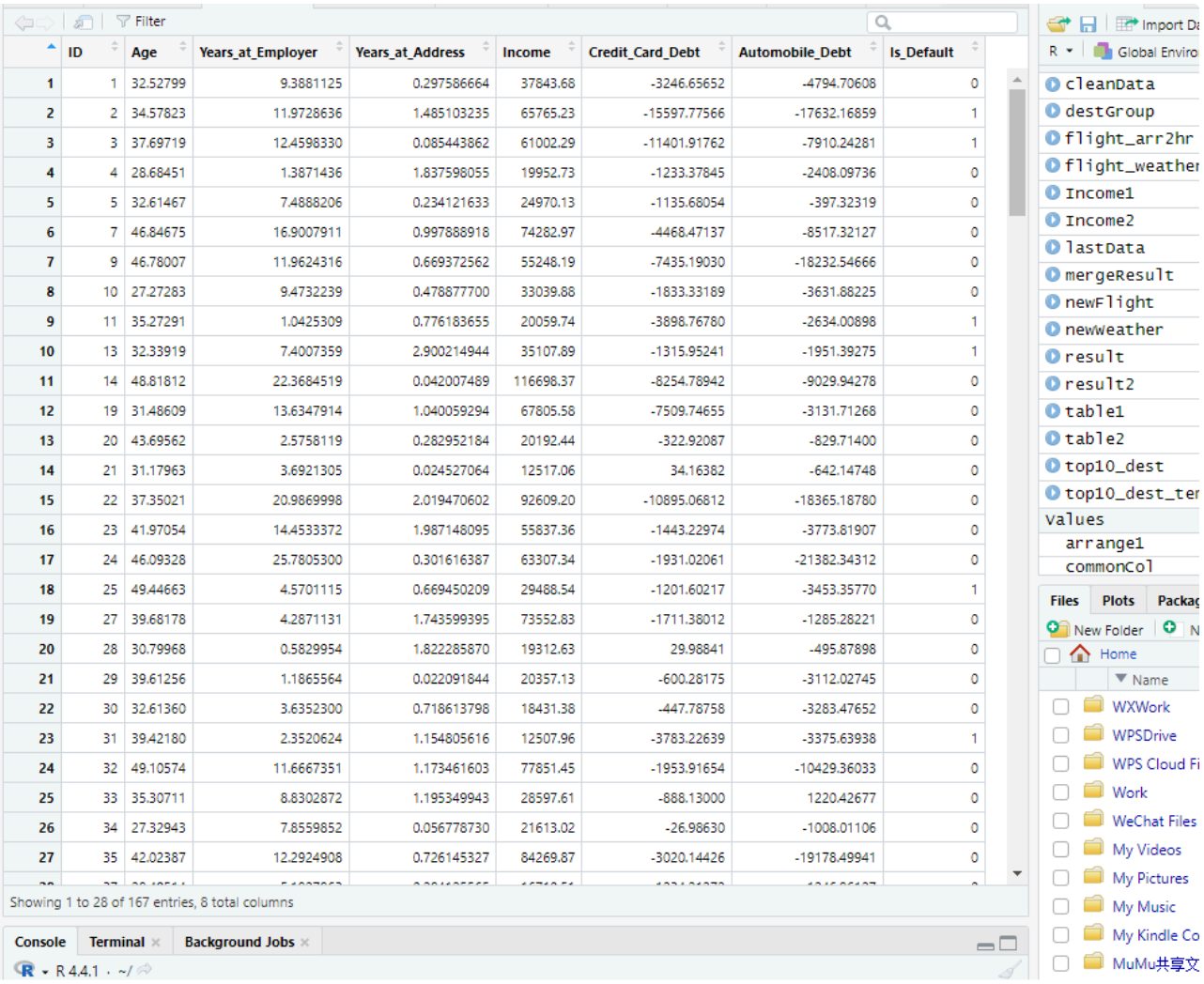
```R
loseData = colSums(is.na(mergeResult))

cleanData = drop_na(mergeResult, "Credit_Card_Debt", "Automobile_Debt", "Is_Default")
```

```
> colSums(is.na(mergeResult))
                ID                Age Years_at_Employer   Years_at_Address             Income
                 0                  0                  0                  0                  0
 Credit_Card_Debt    Automobile_Debt         Is_Default
                11                 11                 11
> loseData = colSums(is.na(mergeResult))
>
> loseData = colSums(is.na(mergeResult))
>
>
```

8. 第八问

```R
removeData = boxplot.stats(cleanData$Income)$out
lastData = cleanData %>% filter(!Income %in% removeData)
```





	ID	Age	Years_at_Employer	Years_at_Address	Income	Credit_Card_Debt	Automobile_Debt	Is_Default
1	1	32.52799	9.3881125	0.297586664	37843.68	-3246.65652	-4794.70608	0
2	2	34.57823	11.9728636	1.485103235	65765.23	-15597.77566	-17632.16859	1
3	3	37.69719	12.4598330	0.085443862	61002.29	-11401.91762	-7910.24281	1
4	4	28.68451	1.3871436	1.837598055	19952.73	-1233.37845	-2408.09736	0
5	5	32.61467	7.4888206	0.234121633	24970.13	-1135.68054	-397.32319	0
6	7	46.84675	16.9007911	0.997888918	74282.97	-4468.47137	-8517.32127	0
7	9	46.78007	11.9624316	0.669372562	55248.19	-7435.19030	-18232.54666	0
8	10	27.27283	9.4732239	0.478877700	33039.88	-1833.33189	-3631.88225	0
9	11	35.27291	1.0425309	0.776183655	20059.74	-3898.76780	-2634.00898	1
10	13	32.33919	7.4007359	2.900214944	35107.89	-1315.95241	-1951.39275	1
11	14	48.81812	22.3684519	0.042007489	116698.37	-8254.78942	-9029.94278	0
12	19	31.48609	13.6347914	1.040059294	67805.58	-7509.74655	-3131.71268	0
13	20	43.69562	2.5758119	0.282952184	20192.44	-322.92087	-829.71400	0
14	21	31.17963	3.6921305	0.024527064	12517.06	34.16382	-642.14748	0
15	22	37.35021	20.9869998	2.019470602	92609.20	-10895.06812	-18365.18780	0
16	23	41.97054	14.4533372	1.987148095	55837.36	-1443.22974	-3773.81907	0
17	24	46.09328	25.7805300	0.301616387	63307.34	-1931.02061	-21382.34312	0
18	25	49.44663	4.5701115	0.669450209	29488.54	-1201.60217	-3453.35770	1
19	27	39.68178	4.2871131	1.743599395	73552.83	-1711.38012	-1285.28221	0
20	28	30.79968	0.5829954	1.822285870	19312.63	29.98841	-495.87898	0
21	29	39.61256	1.1865564	0.022091844	20357.13	-600.28175	-3112.02745	0
22	30	32.61360	3.6352300	0.718613798	18431.38	-447.78758	-3283.47652	0
23	31	39.42180	2.3520624	1.154805616	12507.96	-3783.22639	-3375.63938	1
24	32	49.10574	11.6667351	1.173461603	77851.45	-1953.91654	-10429.36033	0
25	33	35.30711	8.8302872	1.195349943	28597.61	-888.13000	1220.42677	0
26	34	27.32943	7.8559852	0.056778730	21613.02	-26.98630	-1008.01106	0
27	35	42.02387	12.2924908	0.726145327	84269.87	-3020.14426	-19178.49941	0

Showing 1 to 28 of 167 entries, 8 total columns

Console Terminal Background Jobs

R 4.4.1 · ~/

cleanData
destGroup
flight_arr2hr
flight_weather
Income1
Income2
lastData
mergeResult
newFlight
newweather
result
result2
table1
table2
top10_dest
top10_dest_ter
values
arrange1
commonCol

Files Plots Packag

New Folder

Home

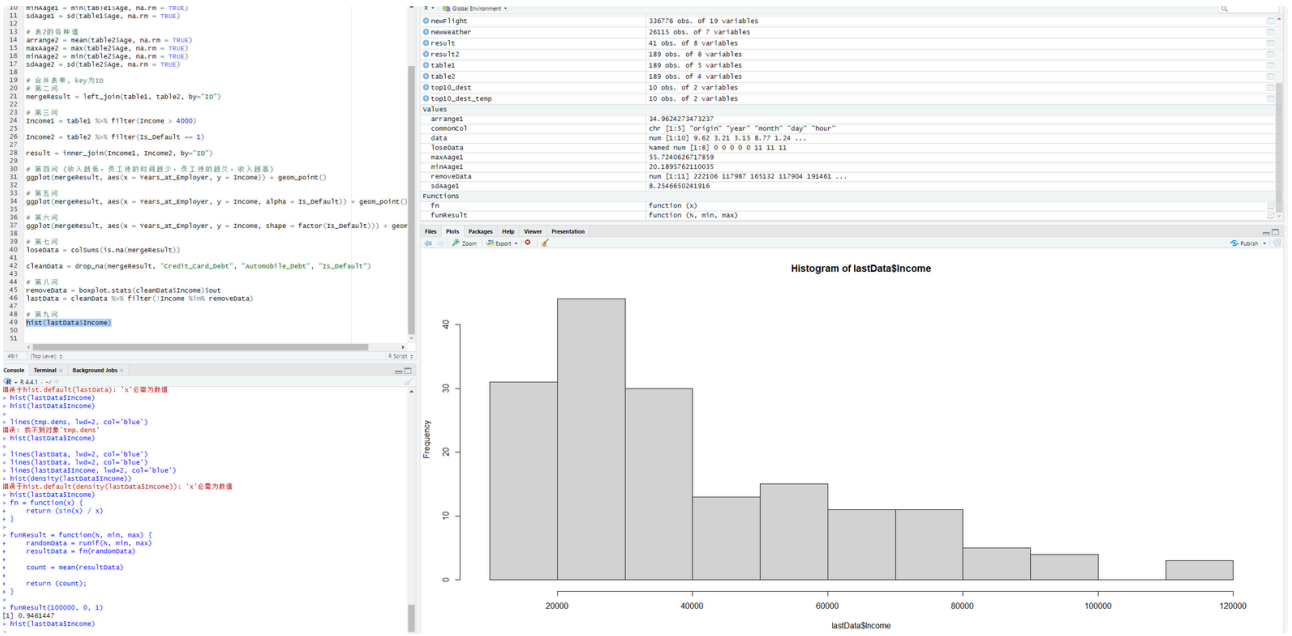
Name

WXWork
WPSDrive
WPS Cloud Fi
Work
WeChat Files
My Videos
My Pictures
My Music
My Kindle Co
MuMu共享文

9. 第九问

R

```
hist(lastData$Income)
```



第三问

```
> hist(lastData$Income)
> fn = function(x) {
+     return (sin(x) / x)
+ }
>
> funResult = function(N, min, max) {
+     randomData = runif(N, min, max)
+     resultData = fn(randomData)
+
+     count = mean(resultData)
+
+     return (count);
+ }
>
> funResult(100000, 0, 1)
[1] 0.9461447
>
```