1

a. readfile rg_train.csv to R. use stringsAsFactors=F (make it a practice)
b. break column age_band into two columns a1 and a2(use - as separator with separate function from package tidyr)
c. Replace value 71+ in a1 with 71
d. convert a1 and a2 to numeric types
e. create a column a3 which takes values using following logic . if a1=71 then a3=71, else a3=(a1+a2)/2
f. Replace missing values in a3 by average of a3
g. remove columns a1, a2 from data

2

a. from family_income columns remove ',' and '<' symbol
b. separate family income into two columns f1 and f2(use >= as separator)
c. convert f1 and f2 into numeric type
d. when f1 is 4000, put 4000 in f2
e. when f2 is 35000 put 35000 in f1
f. create column f3 which is (f1+f2)/2
g. replace missing values in f3 with its mean
h. remove columns f1,f2 from the data

3

a. load data hflights from package hflights
b. Find out for each month count of flights which each unique carrier had
c. Find out which carrier had the highest number of flights for each month
d. Find out , considering the entire year which plane had the highest number of flights for each carrier (planes are identified by tailnum)

4

a. Find out the difference between airtime and actual elapsed time for each observation (time spent on land)
b. Find out the average land time for each month
c. Find out how the average land time changed month over month (make use of difference from lag)

5

use following code to create data

```
set.seed(2)
library(dplyr)
data=data.frame(week=sample(1:4,40,replace = T),
                letr=sample(c("a","b","c","d","e","f","g"),40,replace = T))
data=data %>%
  arrange(week) %>%
```

```
  distinct(week,letr, .keep_all= TRUE) %>%
  mutate(id=1:n())
```

create a new column called output . it takes value “new” if the correponding letr for the observation did not appear in the immediate previous week, otherwise it takes value “repeated”. Here are few initial steps to help you

a. sort the data by (letr,week)
b. group by letr
c. create a new column which takes value week-lag(week)
d. use this new column and any other information to create the output column as stated above.