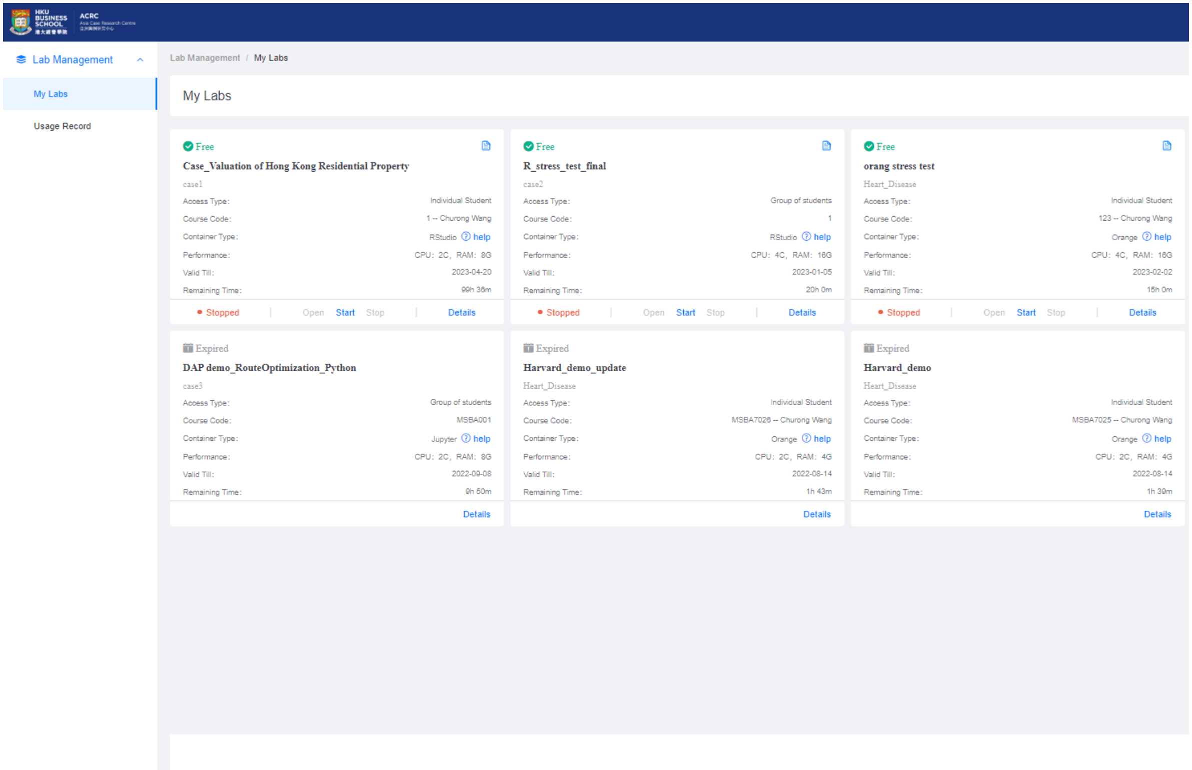
Data Analytics Platform Operation Manual

Student

After lecturers subscribe to a case, students will receive a notification email including information on the link to the Data Analytics Platform, a username (email address of the user), and a temporary password. Students could click on the website link http://47.243.52.252/hku-dap-client/#/Signin to enter the Data Analytics Platform. After entering the email addresses and temporary password, students will be notified to change the temporary password to their password (8-20 digits). Log in to the Platform with the user's email address and password.

My Labs



My Labs

Free
Case_Valuation of Hong Kong Residential Property
case1
Access Type: Individual Student
Course Code: 1 -- Churong Wang
Container Type: RStudio ⓘ help
Performance: CPU: 2C, RAM: 8G
Valid Till: 2023-04-20
Remaining Time: 99h 36m
● Stopped | Open Start Stop | Details

Free
R_stress_test_final
case2
Access Type: Group of students
Course Code: 1
Container Type: RStudio ⓘ help
Performance: CPU: 4C, RAM: 16G
Valid Till: 2023-01-05
Remaining Time: 20h 0m
● Stopped | Open Start Stop | Details

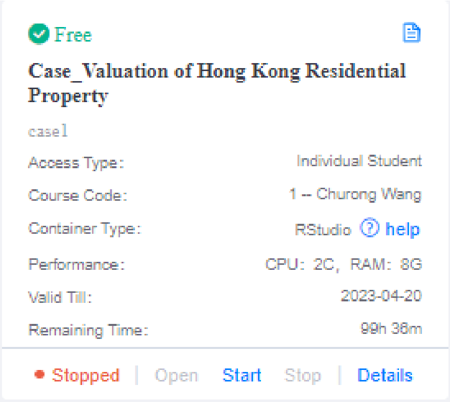
Free
orang stress test
Heart_Disease
Access Type: Individual Student
Course Code: 123 -- Churong Wang
Container Type: Orange ⓘ help
Performance: CPU: 4C, RAM: 16G
Valid Till: 2023-02-02
Remaining Time: 15h 0m
● Stopped | Open Start Stop | Details

Expired
DAP demo_RouteOptimization_Python
case3
Access Type: Group of students
Course Code: MSBA001
Container Type: Jupyter ⓘ help
Performance: CPU: 2C, RAM: 8G
Valid Till: 2022-09-08
Remaining Time: 9h 50m
Details

Expired
Harvard_demo_update
Heart_Disease
Access Type: Individual Student
Course Code: MSBA7026 -- Churong Wang
Container Type: Orange ⓘ help
Performance: CPU: 2C, RAM: 4G
Valid Till: 2022-08-14
Remaining Time: 1h 43m
Details

Expired
Harvard_demo
Heart_Disease
Access Type: Individual Student
Course Code: MSBA7025 -- Churong Wang
Container Type: Orange ⓘ help
Performance: CPU: 2C, RAM: 4G
Valid Till: 2022-08-14
Remaining Time: 1h 39m
Details

Screenshot of the user interface of the Data Analytics Platform



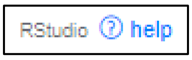
Free
Case_Valuation of Hong Kong Residential Property
case1
Access Type: Individual Student
Course Code: 1 -- Churong Wang
Container Type: RStudio ⓘ help
Performance: CPU: 2C, RAM: 8G
Valid Till: 2023-04-20
Remaining Time: 99h 36m
● Stopped | Open Start Stop | Details

Screenshot of the Labs that students are going to use

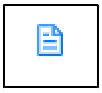
Instructions for using Rstudio

'My Labs' show all the labs that teachers subscribed to for students.

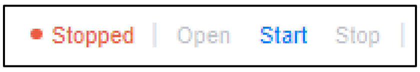
1. For detailed instructions on using the virtual environments, click 'help'.



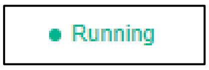
2. For instructions on the case, click the upper right icon.



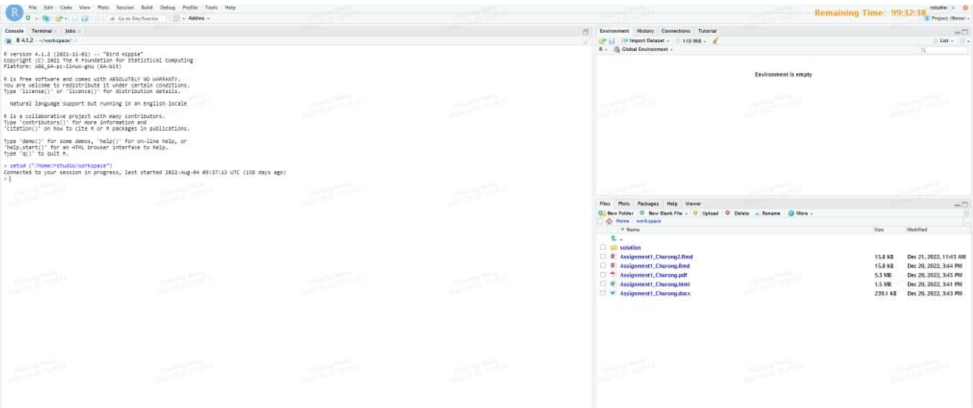
3. To start, open or stop the labs, click the buttons at the bottom of each lab. The left bottom icon shows the status of the lab. Wait for around 15 seconds for the lab to load.



When the status icon turns to "Running," click "Open" to access the virtual environment of Rstudio.



The system will then automatically jump to the virtual environment of Rstudio's user interface. "Remaining Time" shows the time limit of your lab usage. If you use up the lab time, you will need to email the lecturer or the teaching assistant to renew the lab. The watermark shows the user name and the access time of users.

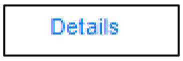


```
R version 4.1.2 (2021-11-01) -- "Bird Hippie"
Copyright (C) 2021 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> setwd ("/home/rstudio/workspace")
Connected to your session in progress, last started 2022-Aug-04 09:37:13 UTC (138 days ago)
> 
```

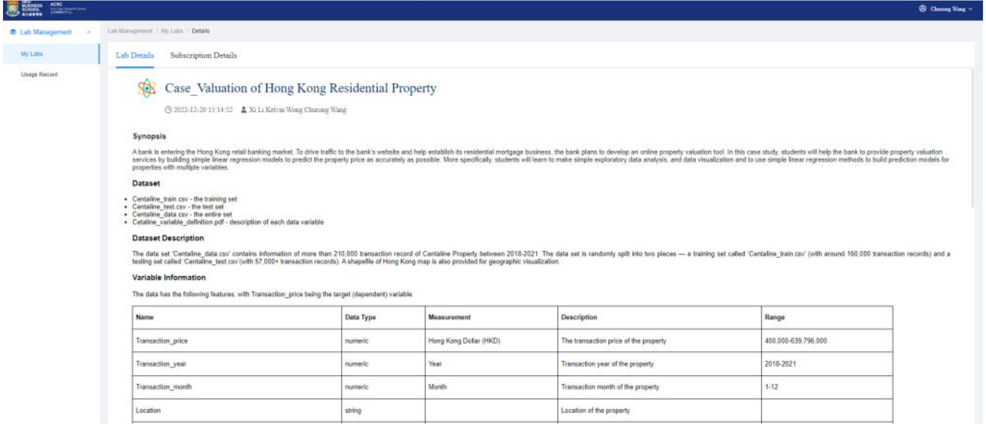
Environment is empty

Name	Size	Modified
..		
solution		
Assignment1_Churong2.Rmd	15.8 KB	Dec 21, 2022, 11:43 AM
Assignment1_Churong.Rmd	15.8 KB	Dec 20, 2022, 3:44 PM
Assignment1_Churong.pdf	5.3 MB	Dec 20, 2022, 3:45 PM
Assignment1_Churong.html	1.5 MB	Dec 20, 2022, 3:41 PM
Assignment1_Churong.docx	239.1 KB	Dec 20, 2022, 3:43 PM

4. Students can view detailed information about each lab by clicking 'Details.'



After entering the "Lab Details" page, you can find more info on the lab.



Case_Valuation of Hong Kong Residential Property

2022-12-20 15:14:52 Xi Li Kelvin Wong Churong Wang

Synopsis

A bank is entering the Hong Kong retail banking market. To drive traffic to the bank's website and help establish its residential mortgage business, the bank plans to develop an online property valuation tool. In this case study, students will help the bank to provide property valuation services by building simple linear regression models to predict the property price as accurately as possible. More specifically, students will learn to make simple exploratory data analysis, and data visualization and to use simple linear regression methods to build prediction models for properties with multiple variables.

Dataset

- Centaline_train.csv - the training set
- Centaline_test.csv - the test set
- Centaline_data.csv - the entire set
- Cetaline_variable_definition.pdf - description of each data variable

Dataset Description

The data set 'Centaline_data.csv' contains information of more than 210,000 transaction record of Centaline Property between 2018-2021. The data set is randomly split into two pieces — a training set called 'Centaline_train.csv' (with around 160,000 transaction records) and a testing set called 'Centaline_test.csv'(with 57,000+ transaction records). A shapefile of Hong Kong map is also provided for geographic visualization.

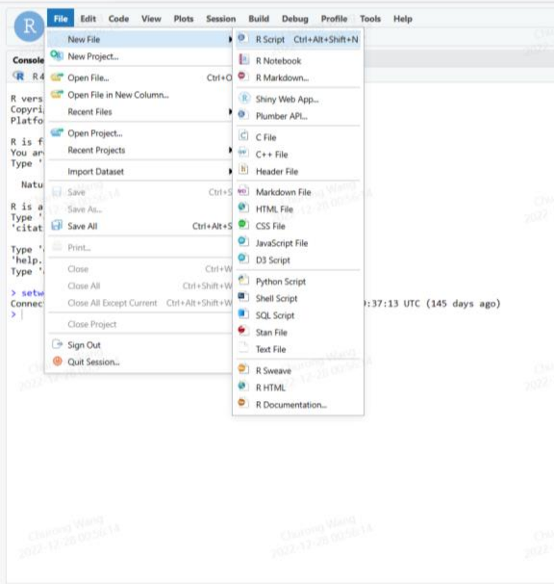
Variable Information

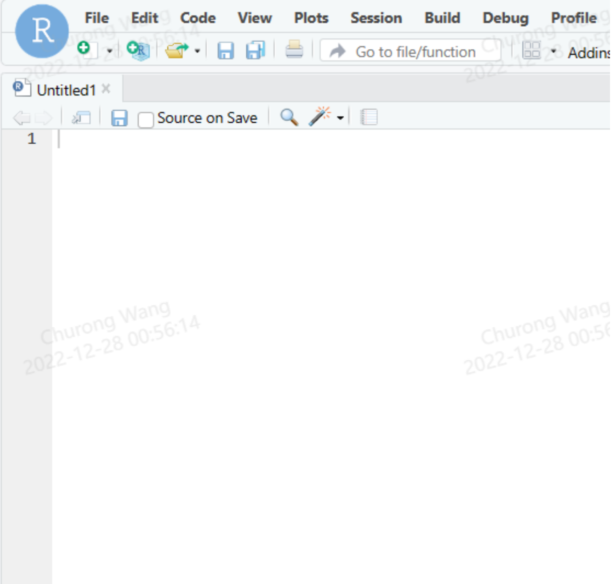
The data has the following features, with Transaction_price being the target (dependent) variable.

Name	Data Type	Measurement	Description	Range
Transaction_price	numeric	Hong Kong Dollar (HKD)	The transaction price of the property	400,000-639,796,000
Transaction_year	numeric	Year	Transaction year of the property	2018-2021
Transaction_month	numeric	Month	Transaction month of the property	1-12
Location	string		Location of the property	

Open a R script file

To create a new R script file, click ‘File>New File > R script’.

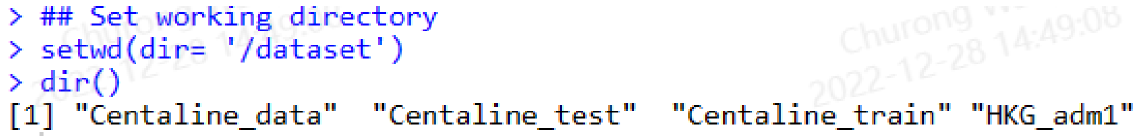




Dataset Access

The system creates the dataset folder by default, and all datasets are stored under the dataset folder. The following codes are examples to access the datasets:

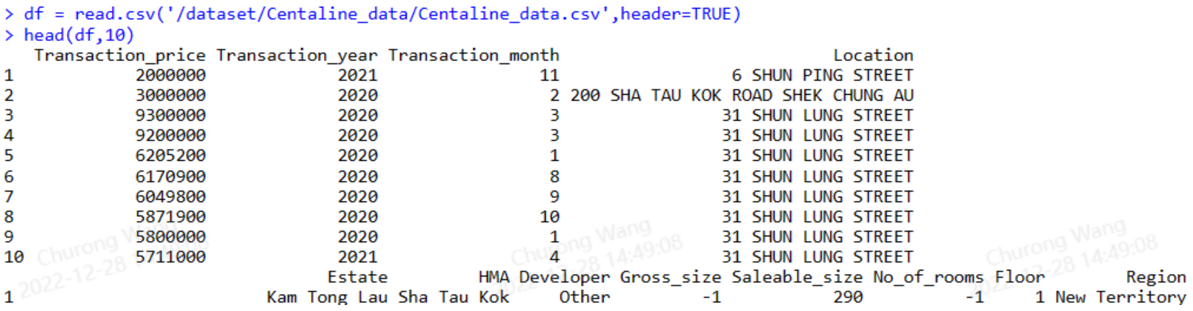
```
## Set working directory
setwd(dir= '/dataset')
dir()
```



```
> ## Set working directory
> setwd(dir= '/dataset')
> dir()
[1] "Centaline_data"  "Centaline_test"  "Centaline_train" "HKG_adm1"
```

You can see there are some datasets stored under the directory “/dataset”. Select your desired dataset from the directory:

```
## load dataset
df = read.csv('/dataset/Centaline_data/Centaline_data.csv',header=TRUE)
head(df,10)
```

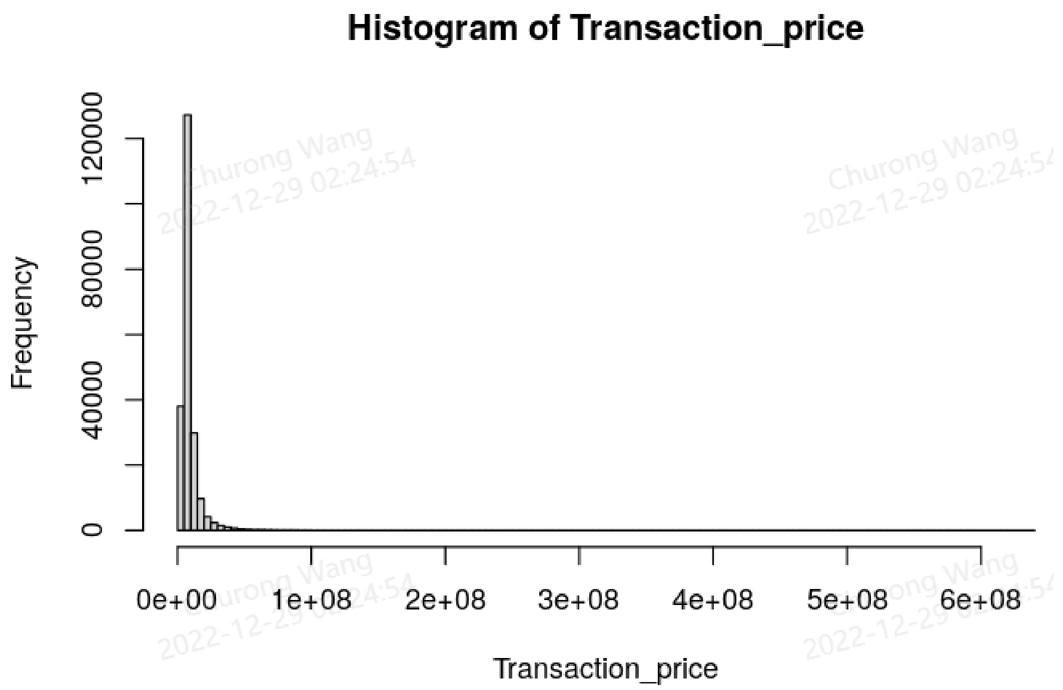


```
> df = read.csv('/dataset/Centaline_data/Centaline_data.csv',header=TRUE)
> head(df,10)
   Transaction_price Transaction_year Transaction_month                         Location
1            2000000             2021                11               6 SHUN PING STREET
2            3000000             2020                 2 200 SHA TAU KOK ROAD SHEK CHUNG AU
3            9300000             2020                 3               31 SHUN LUNG STREET
4            9200000             2020                 3               31 SHUN LUNG STREET
5            6205200             2020                 1               31 SHUN LUNG STREET
6            6170900             2020                 8               31 SHUN LUNG STREET
7            6049800             2020                 9               31 SHUN LUNG STREET
8            5871900             2020                10               31 SHUN LUNG STREET
9            5800000             2020                 1               31 SHUN LUNG STREET
10           5711000             2021                 4               31 SHUN LUNG STREET
                                Estate         HMA Developer Gross_size Saleable_size No_of_rooms Floor        Region
1                         Kam Tong Lau Sha Tau Kok     Other         -1           290          -1     1 New Territory
```

Exploratory Data Analysis (EDA)

Explore the distribution of variables through simple visualizations.

```
## Distribution of Transaction_Price
with(df, hist(Transaction_price,breaks=100))
```



Linear Regression

Linear regression is a statistical model that analyzes the relationship between a response variable (often called y) and one or more variables and their interactions (often called x or explanatory variables). Linear regression can be calculated in R with the command lm(). In the next example, use this command to calculate the property price based on relevant variables. Select a few variables that influence the property price the most. Use transaction_price as the response variable and the other variables as explanatory variables.

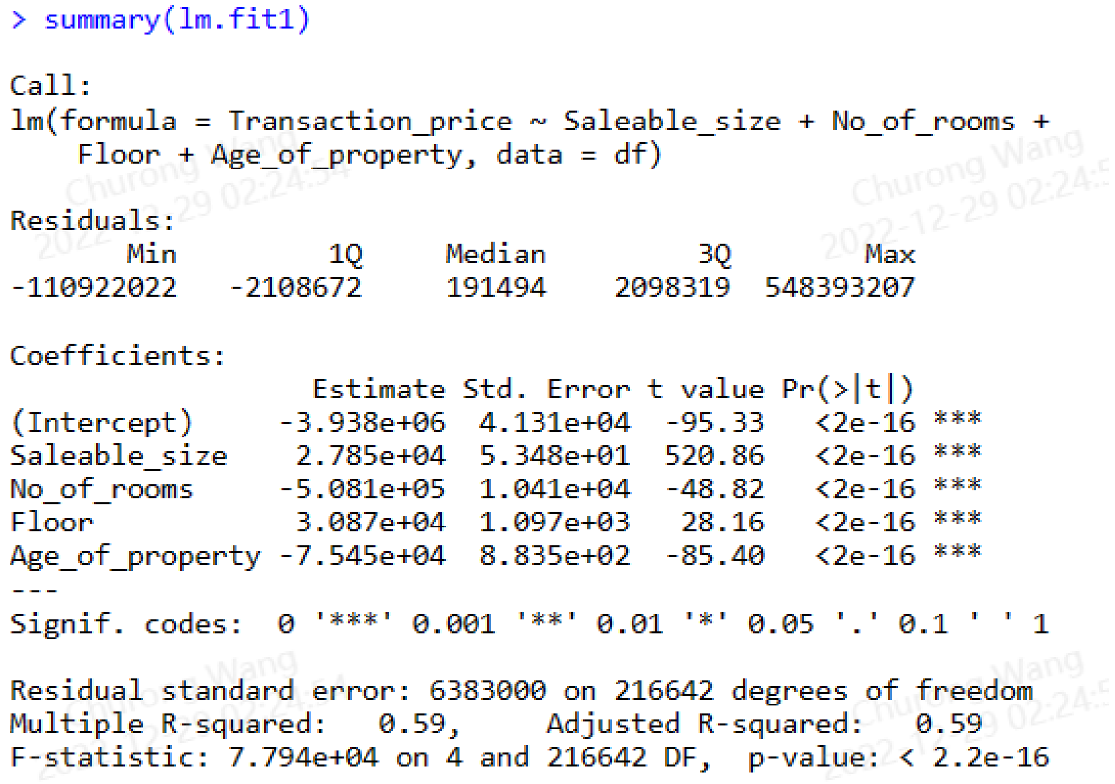
The lm command takes the variables in the format:

lm([target] ~ [predictor / features], data = [data source])

Sample code:

```
lm.fit1=lm(Transaction_price ~
             Saleable_size +
             No_of_rooms +
             Floor +
             Age_of_property,
           data = df)
summary(lm.fit1)
```

Check out the performance of the model with the summary() function.

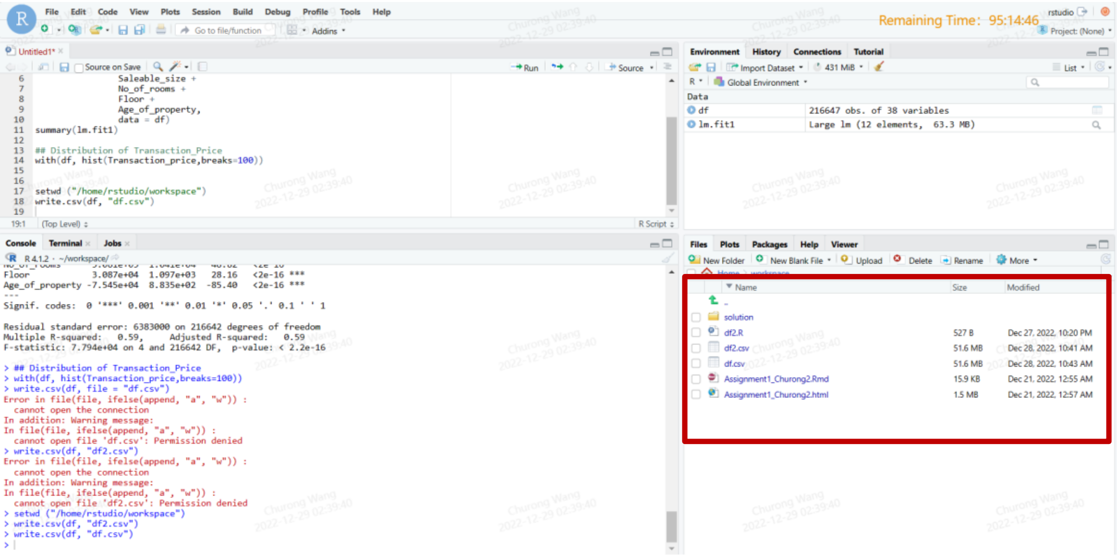


```
> summary(lm.fit1)

Call:
lm(formula = Transaction_price ~ Saleable_size + No_of_rooms +
    Floor + Age_of_property, data = df)

Residuals:
       Min         1Q     Median         3Q        Max
-110922022   -2108672     191494    2098319  548393207

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -3.938e+06  4.131e+04  -95.33   <2e-16 ***
Saleable_size    2.785e+04  5.348e+01  520.86   <2e-16 ***
No_of_rooms     -5.081e+05  1.041e+04  -48.82   <2e-16 ***
Floor            3.087e+04  1.097e+03   28.16   <2e-16 ***
Age_of_property -7.545e+04  8.835e+02  -85.40   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6383000 on 216642 degrees of freedom
Multiple R-squared:   0.59,     Adjusted R-squared:   0.59
F-statistic: 7.794e+04 on 4 and 216642 DF,  p-value: < 2.2e-16
```

Dataset Saving

To save a newly created dataset as a CSV file, use the following sample codes to save it in the 'workspace.' Then, you may find your dataset under "workspace" in the bottom right section of the user interface.

```
## Save dataset to workspace
setwd ("/home/rstudio/workspace")
write.csv(df, file = "df.csv")
```



```
                    Saleable_size +
                    No_of_rooms +
                    Floor +
                    Age_of_property,
                    data = df)
summary(lm.fit1)

## Distribution of Transaction_Price
with(df, hist(Transaction_price,breaks=100))


setwd ("/home/rstudio/workspace")
write.csv(df, "df.csv")
```

```
Floor            3.087e+04  1.097e+03   28.16   <2e-16 ***
Age_of_property -7.545e+04  8.835e+02  -85.40   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6383000 on 216642 degrees of freedom
Multiple R-squared:   0.59,     Adjusted R-squared:   0.59
F-statistic: 7.794e+04 on 4 and 216642 DF,  p-value: < 2.2e-16

> ## Distribution of Transaction_Price
> with(df, hist(Transaction_price,breaks=100))
> write.csv(df, file = "df.csv")
Error in file(file, ifelse(append, "a", "w")) :
  cannot open the connection
In addition: Warning message:
In file(file, ifelse(append, "a", "w")) :
  cannot open file 'df.csv': Permission denied
> write.csv(df, "df2.csv")
Error in file(file, ifelse(append, "a", "w")) :
  cannot open the connection
In addition: Warning message:
In file(file, ifelse(append, "a", "w")) :
  cannot open file 'df2.csv': Permission denied
> setwd ("/home/rstudio/workspace")
> write.csv(df, "df2.csv")
> write.csv(df, "df.csv")
> 
```

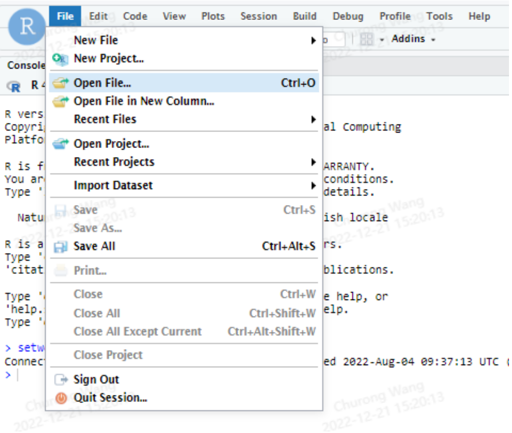
Data

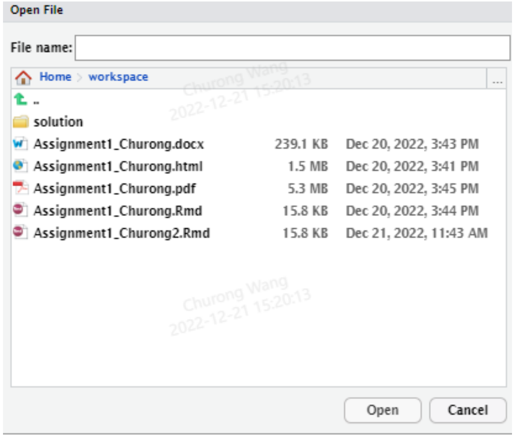
df	216647 obs. of 38 variables
lm.fit1	Large lm (12 elements, 63.3 MB)

Name	Size	Modified
..		
solution		
df2.R	527 B	Dec 27, 2022, 10:20 PM
df2.csv	51.6 MB	Dec 28, 2022, 10:41 AM
df.csv	51.6 MB	Dec 28, 2022, 10:43 AM
Assignment1_Churong2.Rmd	15.9 KB	Dec 21, 2022, 12:55 AM
Assignment1_Churong2.html	1.5 MB	Dec 21, 2022, 12:57 AM

File Access

All files are stored in a folder called “workspace.” Access the “workspace” by clicking “File > Open File > workspace.”

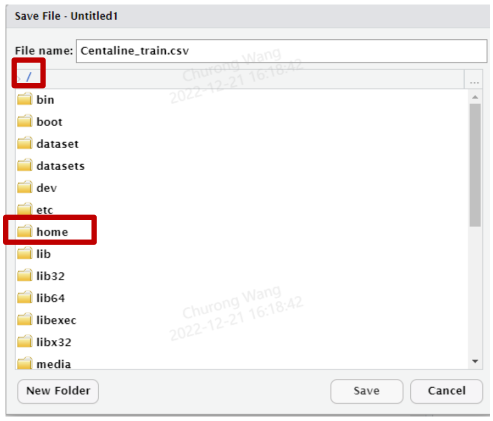




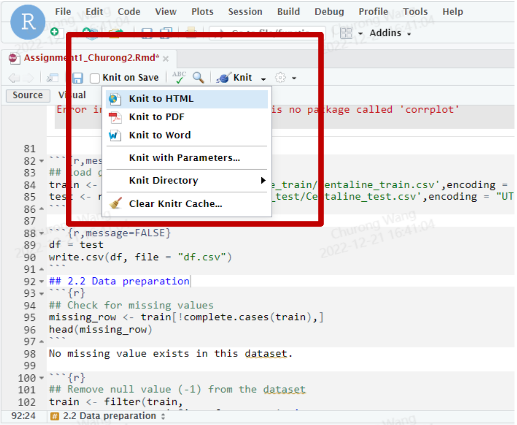
..		
solution		
Assignment1_Churong.docx	239.1 KB	Dec 20, 2022, 3:43 PM
Assignment1_Churong.html	1.5 MB	Dec 20, 2022, 3:41 PM
Assignment1_Churong.pdf	5.3 MB	Dec 20, 2022, 3:45 PM
Assignment1_Churong.Rmd	15.8 KB	Dec 20, 2022, 3:44 PM
Assignment1_Churong2.Rmd	15.8 KB	Dec 21, 2022, 11:43 AM

File Saving

All files must be saved in the “workspace” for future access. Save R files to “workspace” by clicking “File > Save” or “File > Save As…” and browse to the directory “/ > home > rstudio > workspace.” Export R codes to HTML/PDF/Word format by clicking ‘Knit > Knit to …’.



File name: Centaline_train.csv

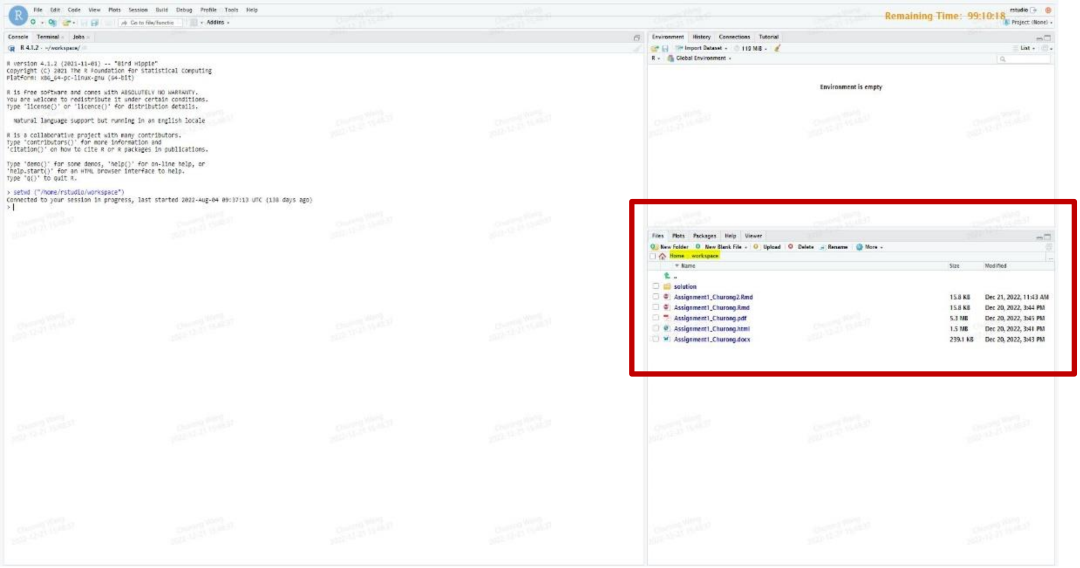


~~~
Er or i                          is no pa kage called 'corrplot'

```{r,mes
## load
train <-                      _train/ entaline_train.csv',encoding =
test <- r                     _test/Ce taline_test.csv',encoding = "UT
~~~

```{r,message=FALSE}
df = test
write.csv(df, file = "df.csv")
```
2.2 Data preparation
```{r}
## Check for missing values
missing_row <- train[!complete.cases(train),]
head(missing_row)
```
No missing value exists in this dataset.

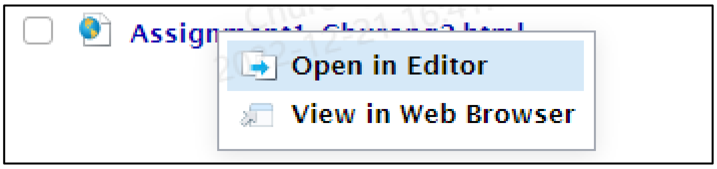
```{r}
## Remove null value (-1) from the dataset
train <- filter(train,
```

To check files other than R format (e.g., PDF/HTML/Word format), check out files on the bottom right of the user interface. The “Files” section displays all the files under the folder ‘workspace.’ Double-click the files to see them on a pop-up web page. For HTML format files, click it and select “View in Web Browser” to open the HTML file in a new web page.



```
R version 4.1.2 (2021-11-01) -- "Bird Hippie"
Copyright (C) 2021 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> setwd ("/home/rstudio/workspace")
Connected to your session in progress, last started 2022-Aug-04 09:37:13 UTC (138 days ago)
> 
```

Name	Size	Modified
..		
solution		
Assignment1_Churong2.Rmd	15.8 KB	Dec 21, 2022, 11:43 AM
Assignment1_Churong.Rmd	15.8 KB	Dec 20, 2022, 3:44 PM
Assignment1_Churong.pdf	5.3 MB	Dec 20, 2022, 3:45 PM
Assignment1_Churong.html	1.5 MB	Dec 20, 2022, 3:41 PM
Assignment1_Churong.docx	239.1 KB	Dec 20, 2022, 3:43 PM



Final Note:

1. Please save all files into the folder “workspace,” and ensure that your UID (student ID) and name are contained in the file name when saving your files. These two steps are very important because if you don't follow these instructions, your assignment may not be visible/identifiable to course instructors to grade your assignment.
2. Labs will automatically close after 30 mins of inactivity. Please always remember to save ALL your working files to “workspace” so you can retrieve your previous work in the future.
3. To save your lab time, remember to click ‘Stop’ to close the labs when you do not use the lab service.