

前處理

註記：

- 使用的資料集為：「taipei_rent.csv」時間為0518. 15:00 從github pull
- 目前沒有看到距離捷運站的距離
- 主題的對象：一般人租屋且以住為主的需求。
- 數值型資料都還沒做區間的分類（討論完確定後再做～）

整理資料：

操作說明	#補原因(為什麼都是NA值) 依據先前討論且該列全部為NA值的欄位刪除，分別為「非都市土地使用分區」「非都市土地使用編定」
程式碼	df <- subset(df, select = -`非都市土地使用分區`) df <- subset(df, select = -`非都市土地使用編定`)
輸出	
結論	

操作說明	確認第一個欄位的狀況，是否有奇怪的類別
程式碼	<pre>distinct(df["鄉鎮市區"]) table(df\$鄉鎮市區, useNA = "ifany") %>% as.data.frame() %>% arrange(desc(Freq)) %>% mutate(Percent = Freq / sum(Freq) * 100) %>% mutate(Percent = round(Percent, 2)) %>% mutate(Percent = paste0(Percent, "%")) %>% rename(鄉鎮市區 = Var1, 數量 = Freq) %>% select(鄉鎮市區, 數量, Percent)</pre>

	<table><tr><th colspan="4">鄉鎮市區 數量 Percent</th></tr><tr><td>1</td><td>中山區</td><td>2515</td><td>17.53%</td></tr><tr><td>2</td><td>士林區</td><td>1692</td><td>11.8%</td></tr><tr><td>3</td><td>萬華區</td><td>1517</td><td>10.58%</td></tr><tr><td>4</td><td>北投區</td><td>1296</td><td>9.04%</td></tr><tr><td>5</td><td>信義區</td><td>1130</td><td>7.88%</td></tr><tr><td>6</td><td>內湖區</td><td>1102</td><td>7.68%</td></tr><tr><td>7</td><td>文山區</td><td>1092</td><td>7.61%</td></tr><tr><td>8</td><td>大同區</td><td>1061</td><td>7.4%</td></tr><tr><td>9</td><td>大安區</td><td>1008</td><td>7.03%</td></tr><tr><td>10</td><td>中正區</td><td>753</td><td>5.25%</td></tr><tr><td>11</td><td>松山區</td><td>723</td><td>5.04%</td></tr><tr><td>12</td><td>南港區</td><td>455</td><td>3.17%</td></tr></table>	鄉鎮市區 數量 Percent				1	中山區	2515	17.53%	2	士林區	1692	11.8%	3	萬華區	1517	10.58%	4	北投區	1296	9.04%	5	信義區	1130	7.88%	6	內湖區	1102	7.68%	7	文山區	1092	7.61%	8	大同區	1061	7.4%	9	大安區	1008	7.03%	10	中正區	753	5.25%	11	松山區	723	5.04%	12	南港區	455	3.17%
鄉鎮市區 數量 Percent																																																					
1	中山區	2515	17.53%																																																		
2	士林區	1692	11.8%																																																		
3	萬華區	1517	10.58%																																																		
4	北投區	1296	9.04%																																																		
5	信義區	1130	7.88%																																																		
6	內湖區	1102	7.68%																																																		
7	文山區	1092	7.61%																																																		
8	大同區	1061	7.4%																																																		
9	大安區	1008	7.03%																																																		
10	中正區	753	5.25%																																																		
11	松山區	723	5.04%																																																		
12	南港區	455	3.17%																																																		
輸出	<table><tr><td colspan="2">鄉鎮市區</td></tr><tr><td>1</td><td>內湖區</td></tr><tr><td>2</td><td>大安區</td></tr><tr><td>3</td><td>中山區</td></tr><tr><td>4</td><td>信義區</td></tr><tr><td>5</td><td>士林區</td></tr><tr><td>6</td><td>松山區</td></tr><tr><td>7</td><td>文山區</td></tr><tr><td>8</td><td>北投區</td></tr><tr><td>9</td><td>萬華區</td></tr><tr><td>10</td><td>南港區</td></tr><tr><td>11</td><td>中正區</td></tr><tr><td>12</td><td>大同區</td></tr></table>	鄉鎮市區		1	內湖區	2	大安區	3	中山區	4	信義區	5	士林區	6	松山區	7	文山區	8	北投區	9	萬華區	10	南港區	11	中正區	12	大同區																										
鄉鎮市區																																																					
1	內湖區																																																				
2	大安區																																																				
3	中山區																																																				
4	信義區																																																				
5	士林區																																																				
6	松山區																																																				
7	文山區																																																				
8	北投區																																																				
9	萬華區																																																				
10	南港區																																																				
11	中正區																																																				
12	大同區																																																				

結論	一切正常，跳過。(後面整理完可以再回來看各自分區的資料量情況(如圖))

操作說明	<p>我認為「主要用途」會刪除大部分不合適的資料，所以從這裡先處理。</p> <p>首先，我先看裡面總共有哪些類別(太多又雜就沒放在結果)；然而，真的太複雜，所以我選擇利用包含的相關字詞內容分成「住宅類」「住商混合」「商業用途」「工業用途」「特殊用途」。</p> <p>分完後，稍微快速瀏覽原始的用途，分成這幾類後每類有多少，而有沒有被分錯或太奇怪的；除此之外，根據我們的主題應該只會留下住宅類和住商混合為主。(這裡有稍微查看每一類的資料量，後續分析是使用保留這兩類的情況！)</p> <p>然而，住宅類中有一項的內容不太符合，為「第八組：社會福利設施—兒童、少年福利機構、防空避難室兼停車空間、第三組：寄宿住宅、機房及水箱、停車空間，機房、第十九組：」，所以刪掉該類別。</p>
程式碼	<pre>df <- df %>% mutate(主要用途 = as.character(主要用途)) # 用文字內容做初步的分類(創建為新的欄位，叫做"主要用途_分類") df <- df %>% mutate(主要用途_分類 = case_when(str_detect(主要用途, "住宅 住家用 集合住宅 多戶住宅 公寓") ~ "住宅類", str_detect(主要用途, "住商 住工 住宅、店舖") ~ "住商混合", str_detect(主要用途, "商業 辦公 事務所 零售業 店舖") ~ "商業用途", str_detect(主要用途, "工業 工廠 廠房 倉儲") ~ "工業用途", str_detect(主要用途, "防空 醫 學 福利 宿舍 交通") ~ "特殊用途", str_trim(主要用途) == "" is.na(主要用途) ~ "未知", TRUE ~ "其他"))</pre>

	<pre> distinct(df["主要用途_分類"]) # 查看每一個分類後內有幾個原始類別 df %>% group_by(主要用途_分類) %>% summarise(原始類別 = paste(unique(主要用途), collapse = ",")) %>% mutate(類別數 = str_count(原始類別, ",") + 1) # 列出所有分類內的原始類別, 確認是否有分錯或太奇怪的 用途清單 <- df %>% group_by(主要用途_分類) %>% summarise(原始用途 = sort(unique(主要用途))) %>% tidyr::unnest(cols = c(原始用途)) print(用途清單, n = Inf) # 看分完後的每個類別有幾筆資料 table(df\$主要用途_分類, useNA = "ifany") %>% as.data.frame() %>% arrange(desc(Freq)) %>% mutate(Percent = Freq / sum(Freq) * 100) %>% mutate(Percent = round(Percent, 2)) %>% mutate(Percent = paste0(Percent, "%")) %>% rename(主要用途_分類 = Var1, 數量 = Freq) %>% select(主要用途_分類, 數量, Percent) # 刪除住宅類我覺得偏奇怪的 df %>% filter(主要用途 == "第八組:社會福利設施—兒童、少年福利機構、防空避難室兼停車空間、第三組:寄宿住宅、機房及水箱、停車空間, 機房、第十九組:") %>% nrow() # 保留只有住宅和住商混合的資料 df <- df %>% filter(`主要用途_分類` %in% c("住宅類", "住商混合")) </pre>
輸出	(這裡呈現比較簡單的結果)

	<p>主要用途_分類 原始類別</p> <p><chr> <chr></p> <p>類別數 <dbl></p> <p>2</p> <p>1 住商混合 "住商用、住工用" 95</p> <p>2 住宅類 "住家用、國民住宅、集合住宅、第八組：社會福利設施-兒童、少年福利機構、防空避難室兼停..." 61</p> <p>3 其他 "見其他登記事項、見使用執照、農舍、停車空間、一般旅館業（B 4）、策略性產業（產品包..." 147</p> <p>4 商業用途 "商業用、日常用品零售業、防空避難室、一般零售業、一般零售業（甲組）、一般零售業（甲..." 22</p> <p>5 工業用途 "工業用、公害輕微之工業（廠房）（不含液化石油氣、汽車改裝業及汽車修理（甲種汽車修理..." 27</p> <p>6 未知 " " 1</p> <p>7 特殊用途 "防空避難室、會客室、宿舍、社會福利設施、防空避難室、停車場、員工宿舍、醫事技術業、..."</p> <p># 列出所有分類內的原始類別，確認是否有分類的主類別</p>
	<p>主要用途_分類 數量 Percent</p> <p>1 住宅類 19008 67.96%</p> <p>2 商業用途 4938 17.65%</p> <p>3 未知 1899 6.79%</p> <p>4 其他 909 3.25%</p> <p>5 住商混合 623 2.23%</p> <p>6 工業用途 477 1.71%</p> <p>7 特殊用途 116 0.41%</p>
結論	<p>刪除後剩下約一萬九千筆資料。</p> <p>這裡因為都只保留住宅和住商的資料，所以這兩欄位最後就刪除了（也就沒有原始的類別「主要用途」和「主要用途_分類」）</p>

操作說明	接著，我選擇處理跟主要用途相關的資料，如「都市土地使用分區」；然而他在資料集中的分類結果有點微妙，所以我選擇刪除這個欄位。
程式碼	<pre>distinct(df["都市土地使用分區"]) table(df\$都市土地使用分區, useNA = "ifany") %>% as.data.frame() %>% arrange(desc(Freq)) %>% mutate(Percent = Freq / sum(Freq) * 100) %>% mutate(Percent = round(Percent, 2)) %>% mutate(Percent = paste0(Percent, "%")) %>% rename(都市土地使用分區 = Var1, 數量 = Freq) %>% select(都市土地使用分區, 數量, Percent) # 我選擇刪除這個欄位</pre>

	<code>df <- subset(df, select = -`都市土地使用分區`)</code>
輸出	<pre> 都市土地使用分區 數量 Percent 1 16114 82.08% 2 住 3517 17.92% </pre>
結論	<p>明明主要用途都是住宅相關, 卻只有部分分區為「住」</p> <p>#補原因(個人認為是因為台北市的發展及過去房屋相關政策有相關)</p>

操作說明	<p>接著, 由於現在的資料已經縮減大部分了, 然而車位這件事, 根據先前討論認為會大幅影響價格這件事, 且同時怕大幅度影響資料的數量, 所以先做處理。</p> <p>先看每個車位類別的資料量, 再做決定。</p>
程式碼	<pre> table(df\$車位類別, useNA = "ifany") %>% as.data.frame() %>% arrange(desc(Freq)) %>% mutate(Percent = Freq / sum(Freq) * 100) %>% mutate(Percent = round(Percent, 2)) %>% mutate(Percent = paste0(Percent, "%")) %>% rename(車位類別 = Var1, 數量 = Freq) %>% select(車位類別, 數量, Percent) # 刪掉所有有車位類別的資料 df <- df %>% filter(`車位類別` %in% c("")) distinct(df["車位面積平方公尺"]) distinct(df["車位總額元"]) df <- subset(df, select = c(-`車位類別`, -`車位面積平方公尺`, -`車位總額元`)) </pre>

輸出	<table><thead><tr><th>車位類別</th><th>數量</th><th>Percent</th></tr></thead><tbody><tr><td>1</td><td>18041</td><td>91.9%</td></tr><tr><td>2 坡道平面</td><td>1238</td><td>6.31%</td></tr><tr><td>3 坡道機械</td><td>106</td><td>0.54%</td></tr><tr><td>4 升降機械</td><td>91</td><td>0.46%</td></tr><tr><td>5 一樓平面</td><td>44</td><td>0.22%</td></tr><tr><td>6 升降平面</td><td>44</td><td>0.22%</td></tr><tr><td>7 塔式車位</td><td>43</td><td>0.22%</td></tr><tr><td>8 其他</td><td>24</td><td>0.12%</td></tr></tbody></table>	車位類別	數量	Percent	1	18041	91.9%	2 坡道平面	1238	6.31%	3 坡道機械	106	0.54%	4 升降機械	91	0.46%	5 一樓平面	44	0.22%	6 升降平面	44	0.22%	7 塔式車位	43	0.22%	8 其他	24	0.12%
車位類別	數量	Percent																										
1	18041	91.9%																										
2 坡道平面	1238	6.31%																										
3 坡道機械	106	0.54%																										
4 升降機械	91	0.46%																										
5 一樓平面	44	0.22%																										
6 升降平面	44	0.22%																										
7 塔式車位	43	0.22%																										
8 其他	24	0.12%																										
結論	根據結果可以發現沒有包含車位的資料仍有一萬八千筆左右，所以直接刪除所有有資料的。																											

操作說明	<p>這部分處理主要建材，只刪除NA值(不多)然後做分類。</p> <p>第一步先做文字上的粗略分類;再根據這個分類結果，因為分布偏激嚴重，決定再根據這個結果再分類(如最後結果)</p> <p>這部分有刪除少量的NA值，而其他部分都沒有被刪除！</p>
程式碼	<pre>distinct(df["主要建材"]) table(df\$主要建材, useNA = "ifany") %>% as.data.frame() %>% arrange(desc(Freq)) %>% mutate(Percent = Freq / sum(Freq) * 100) %>% mutate(Percent = round(Percent, 2)) %>%</pre>

```

mutate(Percent = paste0(Percent, "%")) %>%
rename(車位類別 = Var1, 數量 = Freq) %>%
select(車位類別, 數量, Percent)

df <- df %>%
mutate(主要建材 = case_when(
  str_detect(主要建材, "鋼筋混凝土|RC") ~ "鋼筋混凝土造",
  str_detect(主要建材, "加強磚造") ~ "加強磚造",
  str_detect(主要建材, "鋼骨鋼筋混凝土|鋼骨混凝土") ~ "鋼骨鋼筋混
凝土造",
  str_detect(主要建材, "鋼骨造") ~ "鋼骨造",
  str_detect(主要建材, "磚造") ~ "磚造",
  str_detect(主要建材, "木|竹") ~ "木造",
  str_detect(主要建材, "見使用執照|見其他登記事項") ~
NA_character_,
  str_trim(主要建材) == "" ~ NA_character_,
  TRUE ~ 主要建材
))

# 看看整理後的狀況
df %>%
count(主要建材, sort = TRUE) %>%
mutate(Percent = round(n / sum(n) * 100, 2),
  Percent = paste0(Percent, "%"))

# 刪掉缺失值
df <- df %>%
filter(!is.na(主要建材))

# 再把分出來的類別再縮小類別！
df <- df %>%
mutate(主要建材 = case_when(
  str_detect(主要建材, "鋼筋混凝土造|鋼骨鋼筋混凝土造") ~ "鋼筋混
凝土造類",
  str_detect(主要建材, "加強磚造|磚造|磚石造|土磚石混合造|加強石造|
石造") ~ "加強磚造類",
  str_detect(主要建材, "鋼骨造|鋼骨|鋼構造") ~ "鋼骨類",
  str_detect(主要建材, "木造") ~ "木造類",
  is.na(主要建材) | 主要建材 == "" ~ "其他",
  TRUE ~ "其他"
))

```


	<pre># 看結果狀況 df %>% count(主要建材, sort = TRUE) %>% mutate(Percent = round(n / sum(n) * 100, 2), Percent = paste0(Percent, "%"))</pre>																																																																												
輸出	<table><thead><tr><th></th><th>主要建材</th><th>n</th><th>Percent</th></tr></thead><tbody><tr><td>1</td><td>鋼筋混凝土造</td><td>15969</td><td>88.52%</td></tr><tr><td>2</td><td>加強磚造</td><td>1572</td><td>8.71%</td></tr><tr><td>3</td><td>鋼骨造</td><td>324</td><td>1.8%</td></tr><tr><td>4</td><td><NA></td><td>86</td><td>0.48%</td></tr><tr><td>5</td><td>磚造</td><td>48</td><td>0.27%</td></tr><tr><td>6</td><td>鋼骨鋼筋混凝土造</td><td>23</td><td>0.13%</td></tr><tr><td>7</td><td>磚石造</td><td>6</td><td>0.03%</td></tr><tr><td>8</td><td>土磚石混合造</td><td>4</td><td>0.02%</td></tr><tr><td>9</td><td>木造</td><td>4</td><td>0.02%</td></tr><tr><td>10</td><td>鋼骨</td><td>2</td><td>0.01%</td></tr><tr><td>11</td><td>加強石造</td><td>1</td><td>0.01%</td></tr><tr><td>12</td><td>石造</td><td>1</td><td>0.01%</td></tr><tr><td>13</td><td>鋼構造</td><td>1</td><td>0.01%</td></tr></tbody></table> <table><thead><tr><th></th><th>建材分類</th><th>n</th><th>Percent</th></tr></thead><tbody><tr><td>1</td><td>鋼筋混凝土造類</td><td>15992</td><td>89.07%</td></tr><tr><td>2</td><td>加強磚造類</td><td>1632</td><td>9.09%</td></tr><tr><td>3</td><td>鋼骨類</td><td>327</td><td>1.82%</td></tr><tr><td>4</td><td>木造類</td><td>4</td><td>0.02%</td></tr></tbody></table>		主要建材	n	Percent	1	鋼筋混凝土造	15969	88.52%	2	加強磚造	1572	8.71%	3	鋼骨造	324	1.8%	4	<NA>	86	0.48%	5	磚造	48	0.27%	6	鋼骨鋼筋混凝土造	23	0.13%	7	磚石造	6	0.03%	8	土磚石混合造	4	0.02%	9	木造	4	0.02%	10	鋼骨	2	0.01%	11	加強石造	1	0.01%	12	石造	1	0.01%	13	鋼構造	1	0.01%		建材分類	n	Percent	1	鋼筋混凝土造類	15992	89.07%	2	加強磚造類	1632	9.09%	3	鋼骨類	327	1.82%	4	木造類	4	0.02%
	主要建材	n	Percent																																																																										
1	鋼筋混凝土造	15969	88.52%																																																																										
2	加強磚造	1572	8.71%																																																																										
3	鋼骨造	324	1.8%																																																																										
4	<NA>	86	0.48%																																																																										
5	磚造	48	0.27%																																																																										
6	鋼骨鋼筋混凝土造	23	0.13%																																																																										
7	磚石造	6	0.03%																																																																										
8	土磚石混合造	4	0.02%																																																																										
9	木造	4	0.02%																																																																										
10	鋼骨	2	0.01%																																																																										
11	加強石造	1	0.01%																																																																										
12	石造	1	0.01%																																																																										
13	鋼構造	1	0.01%																																																																										
	建材分類	n	Percent																																																																										
1	鋼筋混凝土造類	15992	89.07%																																																																										
2	加強磚造類	1632	9.09%																																																																										
3	鋼骨類	327	1.82%																																																																										
4	木造類	4	0.02%																																																																										
結論	注意這個類別具有偏態的問題！																																																																												

操作說明	<p>因為這幾個類別都沒有缺失值，我會在每次有做大量資料減少十，觀看一下他們的比例狀況。</p> <p>有無管理組織 有無附傢俱 有無電梯 有無管理員</p>
程式碼	<pre>##### table(df\$有無管理組織, useNA = "ifany") %>% as.data.frame() %>% arrange(desc(Freq)) %>% mutate(Percent = Freq / sum(Freq) * 100) %>% mutate(Percent = round(Percent, 2)) %>% mutate(Percent = paste0(Percent, "%")) %>% rename(有無管理員 = Var1, 數量 = Freq) %>% select(有無管理員, 數量, Percent) ##### table(df\$有無附傢俱, useNA = "ifany") %>% as.data.frame() %>% arrange(desc(Freq)) %>% mutate(Percent = Freq / sum(Freq) * 100) %>% mutate(Percent = round(Percent, 2)) %>% mutate(Percent = paste0(Percent, "%")) %>% rename(有無附傢俱 = Var1, 數量 = Freq) %>% select(有無附傢俱, 數量, Percent) ##### table(df\$有無電梯, useNA = "ifany") %>% as.data.frame() %>% arrange(desc(Freq)) %>% mutate(Percent = Freq / sum(Freq) * 100) %>% mutate(Percent = round(Percent, 2)) %>% mutate(Percent = paste0(Percent, "%")) %>%</pre>

	<pre>rename(有無電梯 = Var1, 數量 = Freq) %>% select(有無電梯, 數量, Percent) ##### table(df\$有無管理員, useNA = "ifany") %>% as.data.frame() %>% arrange(desc(Freq)) %>% mutate(Percent = Freq / sum(Freq) * 100) %>% mutate(Percent = round(Percent, 2)) %>% mutate(Percent = paste0(Percent, "%")) %>% rename(有無管理員 = Var1, 數量 = Freq) %>% select(有無管理員, 數量, Percent) #####</pre>																																																
輸出	<table><tr><th colspan="4">有無管理員 數量 Percent</th></tr><tr><td>1</td><td>無</td><td>7758</td><td>54.09%</td></tr><tr><td>2</td><td>有</td><td>6586</td><td>45.91%</td></tr></table> <table><tr><th colspan="4">有無附傢俱 數量 Percent</th></tr><tr><td>1</td><td>有</td><td>10672</td><td>74.4%</td></tr><tr><td>2</td><td>無</td><td>3672</td><td>25.6%</td></tr></table> <table><tr><th colspan="4">有無電梯 數量 Percent</th></tr><tr><td>1</td><td>無</td><td>7216</td><td>50.31%</td></tr><tr><td>2</td><td>有</td><td>7128</td><td>49.69%</td></tr></table> <table><tr><th colspan="4">有無管理員 數量 Percent</th></tr><tr><td>1</td><td>無</td><td>8596</td><td>59.93%</td></tr><tr><td>2</td><td>有</td><td>5748</td><td>40.07%</td></tr></table>	有無管理員 數量 Percent				1	無	7758	54.09%	2	有	6586	45.91%	有無附傢俱 數量 Percent				1	有	10672	74.4%	2	無	3672	25.6%	有無電梯 數量 Percent				1	無	7216	50.31%	2	有	7128	49.69%	有無管理員 數量 Percent				1	無	8596	59.93%	2	有	5748	40.07%
有無管理員 數量 Percent																																																	
1	無	7758	54.09%																																														
2	有	6586	45.91%																																														
有無附傢俱 數量 Percent																																																	
1	有	10672	74.4%																																														
2	無	3672	25.6%																																														
有無電梯 數量 Percent																																																	
1	無	7216	50.31%																																														
2	有	7128	49.69%																																														
有無管理員 數量 Percent																																																	
1	無	8596	59.93%																																														
2	有	5748	40.07%																																														
結論	注意兩類別的比例即可！																																																

操作說明	「出租型態」這個欄位，我們過去認為他是一個租屋上滿關鍵的部分，畢竟雅房套房就有明顯之別；而分租套房和獨立套房亦不同。
------	--

	選擇刪除約一成比例的缺失值資料。																												
程式碼	<pre># 出租型態 table(df\$出租型態, useNA = "ifany") %>% as.data.frame() %>% arrange(desc(Freq)) %>% mutate(Percent = Freq / sum(Freq) * 100) %>% mutate(Percent = round(Percent, 2)) %>% mutate(Percent = paste0(Percent, "%")) %>% rename(出租型態 = Var1, 數量 = Freq) %>% select(出租型態, 數量, Percent)</pre>																												
輸出	<table><tr><th></th><th>出租型態</th><th>數量</th><th>Percent</th></tr><tr><td>1</td><td>整棟(戶)出租</td><td>8642</td><td>48.13%</td></tr><tr><td>2</td><td>獨立套房</td><td>3126</td><td>17.41%</td></tr><tr><td>3</td><td></td><td>2959</td><td>16.48%</td></tr><tr><td>4</td><td>分租套房</td><td>2454</td><td>13.67%</td></tr><tr><td>5</td><td>分租雅房</td><td>650</td><td>3.62%</td></tr><tr><td>6</td><td>分層出租</td><td>124</td><td>0.69%</td></tr></table>		出租型態	數量	Percent	1	整棟(戶)出租	8642	48.13%	2	獨立套房	3126	17.41%	3		2959	16.48%	4	分租套房	2454	13.67%	5	分租雅房	650	3.62%	6	分層出租	124	0.69%
	出租型態	數量	Percent																										
1	整棟(戶)出租	8642	48.13%																										
2	獨立套房	3126	17.41%																										
3		2959	16.48%																										
4	分租套房	2454	13.67%																										
5	分租雅房	650	3.62%																										
6	分層出租	124	0.69%																										
結論	選擇刪除約一成比例的缺失值資料。																												

操作說明	「租賃住宅服務」關於這部分，選擇保留缺失值，就算沒有這項服務也可以獨立自行完成找房和租房的動作。
程式碼	<pre>table(df\$租賃住宅服務, useNA = "ifany") %>% as.data.frame() %>% arrange(desc(Freq)) %>% mutate(Percent = Freq / sum(Freq) * 100) %>% mutate(Percent = round(Percent, 2)) %>% mutate(Percent = paste0(Percent, "%")) %>%</pre>

	<pre>rename(租賃住宅服務 = Var1, 數量 = Freq) %>% select(租賃住宅服務, 數量, Percent)</pre>
輸出	<pre> 租賃住宅服務 數量 Percent 1 6588 36.69% 2 社會住宅代管 4320 24.06% 3 社會住宅包租轉租 3777 21.04% 4 一般轉租 2768 15.42% 5 一般代管 354 1.97% 6 一般包租 148 0.82%</pre>
結論	空白值要記得補值。

操作說明	<p>再來刪掉一些我覺得不是很重要的欄位內容</p> <p>交易標的: 都是租屋</p> <p>file 編號 備註: 都不重要</p> <p>門牌: 好像沒什麼功能, 對預測而言也沒有</p> <p>土地面積平方公尺: 根據群組的定義, 沒有實質功能</p> <p># 不確定「建物總面積平方公尺」會不會比較有用</p>
程式碼	<pre>distinct(df["交易標的"]) df <- subset(df, select = -`交易標的`) df <- subset(df, select = -`source_file`) df <- subset(df, select = -`編號`) df <- subset(df, select = -`備註`) # 可以暫時保留門牌的欄位 # df <- subset(df, select = -`土地位置建物門牌`) df <- subset(df, select = -`土地面積平方公尺`) # 我覺得型態就已經大致區分了! # 不需要這個樓層數, 反而會影響透天厝等型態的因子 df <- subset(df, select = -`總樓層數`)</pre>

輸出	
結論	

操作說明	把附屬設備內容轉成各個欄位！
程式碼	<pre> 所有設備 <- df\$附屬設備 %>% str_split(",") %>% unlist() %>% trimws() %>% unique() sort(所有設備) 設備列表 <- c("冰箱", "冷氣", "有線電視", "洗衣機", "熱水器", "瓦斯或天然氣", "網路", "電視機") for (item in 設備列表) { df[[item]] <- grepl(item, df\$附屬設備) } df <- df %>% mutate(across(all_of(設備列表), ~ as.integer(.))) df <- subset(df, select = -`附屬設備`) </pre>
輸出	
結論	


操作說明	處理租賃期間轉為天數「租期天數」
程式碼	<pre> df <- df %>% mutate(起始日_raw = str_sub(租賃期間, 1, 7), 結束日_raw = str_sub(租賃期間, 9, 15)) </pre>

	<pre>) convert_minguo_to_date <- function(x) { y <- as.integer(str_sub(x, 1, 3)) + 1911 m <- str_sub(x, 4, 5) d <- str_sub(x, 6, 7) ymd(paste0(y, "-", m, "-", d)) } df <- df %>% mutate(起始日 = convert_minguo_to_date(起始日_raw), 結束日 = convert_minguo_to_date(結束日_raw), 租期天數 = as.numeric(結束日 - 起始日)) summary(df\$租期天數) df <- subset(df, select = `起始日_raw`) df <- subset(df, select = `結束日_raw`) df <- subset(df, select = `租賃期間`) df <- subset(df, select = `起始日`) df <- subset(df, select = `結束日`) </pre>
輸出	
結論	

操作說明	因為我們的對象是一般人的租屋(住)行為, 那關於商業大樓、店面(前面如果是住商混合類型, 如果是用商業應該就會在這裡被歸成店面)以及工廠, 這幾類的資料就刪除了(不多)
程式碼	<pre> table(df\$建物型態, useNA = "ifany") %>% as.data.frame() %>% arrange(desc(Freq)) %>% mutate(Percent = Freq / sum(Freq) * 100) %>% mutate(Percent = round(Percent, 2)) %>% mutate(Percent = paste0(Percent, "%")) %>% rename(建物型態 = Var1, 數量 = Freq) %>% </pre>

	<pre>select(建物型態, 數量, Percent) df <- df %>% filter(!(建物型態 %in% c("辦公商業大樓", "店面(店鋪)", "工廠"))) #####</pre>
輸出	<pre> +-----+-----+-----+ 建物型態 數量 Percent +-----+-----+-----+ 1 公寓(5樓含以下無電梯) 6768 47.18% 2 住宅大樓(11層含以上有電梯) 3802 26.51% 3 華廈(10層含以下有電梯) 3240 22.59% 4 透天厝 476 3.32% 5 其他 58 0.4% +-----+-----+-----+ </pre>
結論	

操作說明	<p>這裡我覺得跟前面的建物型態有點類似。</p> <p>可以討論看看要不要分成透天厝、低樓層、中樓層、高樓層(類別太多)</p> <p>那這裡如果是「全」確定對應的都是透天厝不用擔心！</p>
程式碼	<pre>table(df\$租賃層次, useNA = "ifany") %>% as.data.frame() %>% arrange(desc(Freq)) %>% mutate(Percent = Freq / sum(Freq) * 100) %>% mutate(Percent = round(Percent, 2)) %>% mutate(Percent = paste0(Percent, "%")) %>% rename(租賃層次 = Var1, 數量 = Freq) %>% select(租賃層次, 數量, Percent) # 兩者數量正確！ df %>% filter((租賃層次 == "全" & 建物型態 == "透天厝")) %>% nrow()</pre>

	<pre>df <- df %>% filter(!(租賃層次 %in% c("見其他登記事項")))</pre>
輸出	 <pre>租賃層次 數量 Percent 1 四層 2941 19.71% 2 三層 2584 17.32% 3 二層 2354 15.78% 4 五層 1816 12.17% 5 六層 736 4.93% 6 七層 693 4.65% 7 一層 663 4.44% 8 全 611 4.1% 9 八層 531 3.56% 10 九層 396 2.65% 11 十層 385 2.58% 12 十一層 370 2.48% 13 十二層 327 2.19% 14 十四層 169 1.13% 15 十三層 149 1% 16 十五層 55 0.37% 17 十六層 43 0.29% 18 十七層 37 0.25% 19 十八層 20 0.13% 20 二十層 10 0.07% 21 二十一層 8 0.05% 22 十九層 6 0.04% 23 二十三層 3 0.02% 24 二十四層 3 0.02% 25 地下層 3 0.02% 26 二十二層 2 0.01% 27 地下一層 2 0.01% 28 二十五層 1 0.01% 29 二十六層 1 0.01%</pre>
結論	

操作說明	不確定這裡的欄位在幹嘛；如同程式碼的註解處理方式。
程式碼	<pre> head(df\$租賃筆棟數) df <- df %>% mutate(土地數 = str_extract(租賃筆棟數, "(?<=土地)\\d+") %>% as.integer(), 建物數 = str_extract(租賃筆棟數, "(?<=建物)\\d+") %>% as.integer(), 車位數 = str_extract(租賃筆棟數, "(?<=車位)\\d+") %>% as.integer()) df <- subset(df, select = -`租賃筆棟數`) table(df\$土地數, useNA = "ifany") %>% as.data.frame() %>% arrange(desc(Freq)) %>% mutate(Percent = Freq / sum(Freq) * 100) %>% mutate(Percent = round(Percent, 2)) %>% mutate(Percent = paste0(Percent, "%")) %>% rename(土地數 = Var1, 數量 = Freq) %>% select(土地數, 數量, Percent) table(df\$建物數, useNA = "ifany") %>% as.data.frame() %>% arrange(desc(Freq)) %>% mutate(Percent = Freq / sum(Freq) * 100) %>% mutate(Percent = round(Percent, 2)) %>% mutate(Percent = paste0(Percent, "%")) %>% rename(建物數 = Var1, 數量 = Freq) %>% select(建物數, 數量, Percent) table(df\$車位數, useNA = "ifany") %>% as.data.frame() %>% arrange(desc(Freq)) %>% mutate(Percent = Freq / sum(Freq) * 100) %>% mutate(Percent = round(Percent, 2)) %>% mutate(Percent = paste0(Percent, "%")) %>% rename(車位數 = Var1, 數量 = Freq) %>% </pre>

	<pre> select(車位數, 數量, Percent) # 這裡我選擇刪除車位數, 確定都是零。 df <- subset(df, select = -`車位數`) ## 不確定以下做法是否正確, 所以沒有刪除兩欄位 # 這裡我選擇刪除土地數大於一的資料 # 從土地面積的零的定義, 我覺得大於一的應該就是非單純租房的 df <- df %>% filter(土地數 <= 1) # 這裡我選擇刪除建物數大於一的資料 # 直觀來說我們的對象就是一個想租一間房的人, # 那建物數太多也很怪(反正多的資料也很少) df <- df %>% filter(建物數 <= 1) </pre>
輸出	
結論	

操作說明	因為建築完成年月缺失值不多, 我直接刪除, 然後轉為屋齡
程式碼	<pre> # 刪除租賃年月日, 已經有天數了 df <- subset(df, select = -`租賃年月日`) # 把完成年月轉成屋齡 sum(is.na(df\$建築完成年月)) # 並沒有很多筆, 所以選擇刪除 df <- df %>% filter(!is.na(建築完成年月)) library(dplyr) library(lubridate) df <- df %>% mutate(</pre>

	<pre> 建築完成年月_chr = as.character(建築完成年月), 建築完成年月_chr = str_pad(建築完成年月_chr, width = 7, side = "left", pad = "0"), 建築完成日期 = ymd(paste0(as.integer(substr(建築完成年月_chr, 1, 3)) + 1911, substr(建築完成年月_chr, 4, 6), substr(建築完成年月_chr, 7, 7))), # 距離現在的年數, 無條件進位 屋齡 = ceiling(time_length(interval(建築完成日期, Sys.Date()), "year"))) df <- subset(df, select = -`建築完成年月`) df <- subset(df, select = -`建築完成年月_chr`) df <- subset(df, select = -`建築完成日期`) </pre>
輸出	
結論	