Sprint Review

PDF File Manipulation using PDFtools Library in R

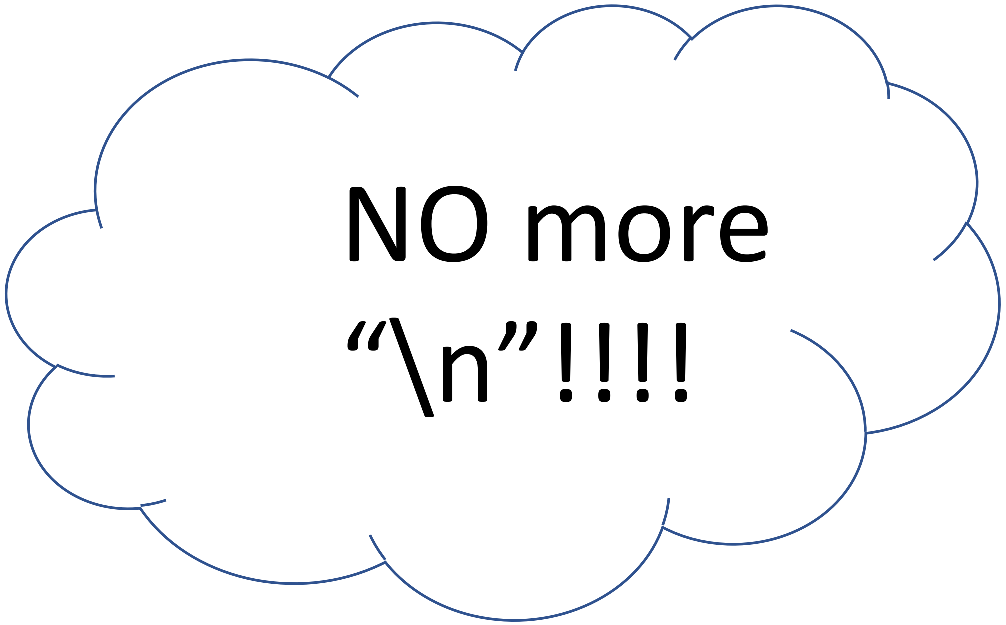
devleague

- library(pdftools) - pdfpages <- pdf_text("HotelReview12.31.17.pdf"). - print(pdfpages)

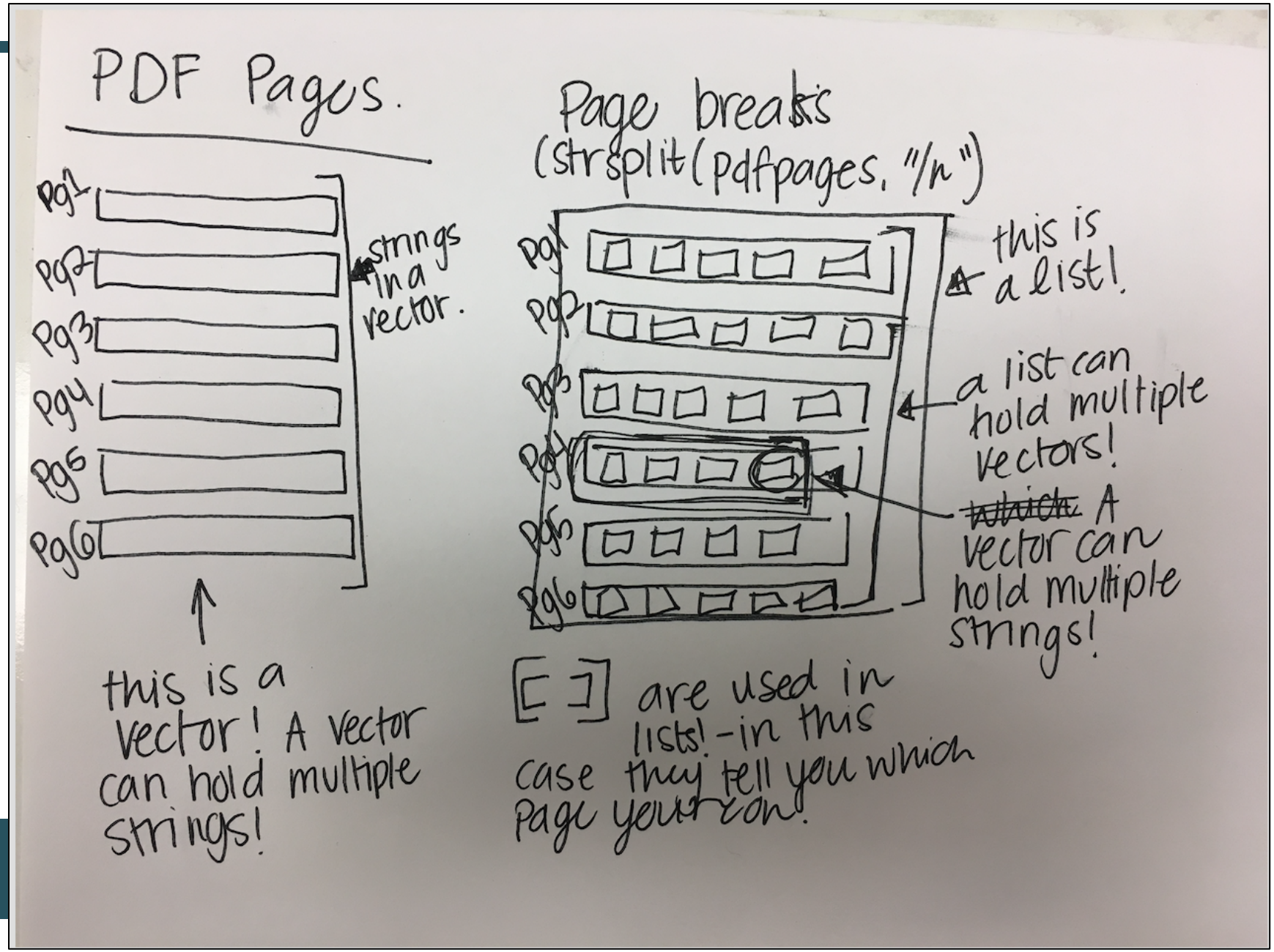
```
 [3] "Performance by Industry Segments\nFor the Week of: December 31, 2017 - January 06,
RevPar by Chain Scale - Current Week
RevPar by Chain Scale - Running 28 Days\n
This Year        Last Year
This Year     Last Year\n                                                    $250
$250\n                                                  $200
$200\n                                                  $150
$150\n                                                  $100
$100\n                                                   $50
$50\n                                                    $0
$0\n                                                          Total United      Luxury      Upper
Midscale      Economy        Independents                        Total United          Luxury
er           Midscale        Economy       Independents\n
Midscale                                                                            States
Midscale\n
Running 28 Days\n                                                        Occ (%)
($)                        Percent Change (%)                                 Occ (%)
RevPAR ($)                                Percent Change (%)\n
2017             2016             2017             2016          Occ              ADR
2017            2016              2017             2016            Occ            ADR
48.7             47.1        124.33           117.50              60.59           55.29
50.4             49.4          119.90           117.96              60.48             58.31
.7\nChain Scale\nLuxury                                              59.8              55.4
```

- pagebreaks = strsplit(pdfpages, "\n") -Print(pagebreaks)

```
 [3] "                                                                              RevPar by Chain Scale - Current Week
RevPar by Chain Scale - Running 28 Days"
 [4] "
This Year      Last Year
This Year     Last Year"
 [5] "                                               $250
$250"
 [6] "                                               $200
$200"
 [7] "                                               $150
$150"
 [8] "                                               $100
$100"
 [9] "                                                $50
$50"
[10] "                                                 $0
$0"
[11] "                                                        Total United       Luxury     Upper Upscale       Upscale          Upper
Midscale      Economy       Independents                        Total United       Luxury     Upper Upscale       Upscale          Upp
er           Midscale      Economy       Independents"
[12] "                                                           States                                                        Midscale
States                                                        Midscale"
[13] "                                                                                                      Current Week
Running 28 Days"
[14] "                                                       Occ (%)                         ADR ($)                         RevPAR ($)
Percent Change (%)                                   Occ (%)                        ADR ($)                         RevPAR ($)
Percent Change (%)"
```



- pdfpage3 <- pagebreaks[[3]]
- What are the [[double brackets]]?
- I’m so glad you asked



devleague

Delete all of the lines with no data

```
datapage = pdfpage3[-(1:10)]
datapage = datapage[-(1:5)]
datapage = datapage[-(2)]
datapage = datapage[-(9)]
datapage = datapage[-(15)]
datapage = datapage[-(21)]
datapage = datapage[-(46:48)]
```

Efficient?

Maybe not.

… All suggestions welcome

datapage ---- before str_extract_all

```
[38] "Philadelphia, PA-NJ                          54.3          46.3       122.52          108.80          66.50          50
.39           17.2          12.6          32.0          51.3          48.0       111.62          108.60          57.21
52.15            6.7           2.8           9.7"
[39] "Phoenix, AZ                                  53.1          51.4       111.30          114.44          59.11          58
.79            3.4          -2.7           0.5          54.9          52.4       106.61          107.46          58.49
56.27            4.8          -0.8           4.0"
[40] "San Diego, CA                                62.9          52.7       127.18          115.54          79.95          60
.84           19.4          10.1          31.4          61.4          59.3       122.90          120.19          75.47
```

datapage = str_extract_all(datapage,"[+-]?([0-9]*[.])?[0-9]+")

```
[[38]]
 [1] "54.3"   "46.3"   "122.52" "108.80" "66.50"  "50.39"  "17.2"   "12.6"   "32.0"   "51.3"   "48.0"   "111.62" "108.60"
[14] "57.21"  "52.15"  "6.7"    "2.8"    "9.7"

[[39]]
 [1] "53.1"   "51.4"   "111.30" "114.44" "59.11"  "58.79"  "3.4"    "-2.7"   "0.5"    "54.9"   "52.4"   "106.61" "107.46"
[14] "58.49"  "56.27"  "4.8"    "-0.8"   "4.0"

[[40]]
 [1] "62.9"   "52.7"   "127.18" "115.54" "79.95"  "60.84"  "19.4"   "10.1"   "31.4"   "61.4"   "59.3"   "122.90" "120.19"
[14] "75.47"  "71.28"  "3.6"    "2.3"    "5.9"
```

Beautiful Organized Strings

“Unlist()” the data = turns all of the strings into one big string.

```
 [1] "48.7"   "47.1"   "124.33" "117.50" "60.59"  "55.29"  "3.6"    "5.8"    "9.6"    "50.4"   "49.4"   "119.90" "117.96"
[14] "60.48"  "58.31"  "2.0"    "1.7"    "3.7"    "59.8"   "55.4"   "399.93" "345.10" "239.02" "191.28" "7.8"    "15.9"
[27] "25.0"   "60.7"   "58.8"   "376.80" "367.71" "228.82" "216.22" "3.3"    "2.5"    "5.8"    "53.6"   "51.2"   "171.81"
[40] "160.67" "92.04"  "82.29"  "4.6"    "6.9"    "11.8"   "55.0"   "53.9"   "164.81" "161.97" "90.60"  "87.23"  "2.1"
[53] "1.8"    "3.9"    "52.2"   "49.9"   "131.38" "123.01" "68.54"  "61.35"  "4.6"    "6.8"    "11.7"   "56.0"   "54.3"
[66] "128.21" "125.63" "71.74"  "68.21"  "3.1"    "2.1"    "5.2"    "46.4"   "44.2"   "102.28" "97.57"  "47.46"  "43.11"
```

Use cbind to create columns in a data.frame!

Hoteldata = cbind.data.frame(split(datapagetest, rep(1:18, times=length(datapagetest)/18)), stringsAsFactors=F)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	48.7	47.1	124.33	117.50	60.59	55.29	3.6	5.8	9.6	50.4	49.4	119.90	117.96	60.48	58.31	2.0	1.7	3.7
2	59.8	55.4	399.93	345.10	239.02	191.28	7.8	15.9	25.0	60.7	58.8	376.80	367.71	228.82	216.22	3.3	2.5	5.8
3	53.6	51.2	171.81	160.67	92.04	82.29	4.6	6.9	11.8	55.0	53.9	164.81	161.97	90.60	87.23	2.1	1.8	3.9
4	52.2	49.9	131.38	123.01	68.54	61.35	4.6	6.8	11.7	56.0	54.3	128.21	125.63	71.74	68.21	3.1	2.1	5.2
5	46.4	44.2	102.28	97.57	47.46	43.11	5.0	4.8	10.1	49.7	48.2	101.22	99.20	50.35	47.82	3.2	2.0	5.3

Name row and columns using

Names() - for columns
row.names() - for rows

AND.....

BEHOLD!!!!!

	CurrentWeek_2017_Occ	CurrentWeek_2016_Occ	CurrentWeek_2017_ADR	CurrentWeek_2016_ADR	CurrentWeek_2017_RevPar	CurrentWeek_2016_RevPar
Total United States	48.7	47.1	124.33	117.50	60.59	55.29
ChainScale_Luxury	59.8	55.4	399.93	345.10	239.02	191.28
ChainScale_Upper Upscale	53.6	51.2	171.81	160.67	92.04	82.29
ChainScale_Upscale	52.2	49.9	131.38	123.01	68.54	61.35
ChainScale_Upper Midscale	46.4	44.2	102.28	97.57	47.46	43.11
ChainScale_Midscale	44.0	42.3	81.57	77.19	35.88	32.65
ChainScale_Economy	46.7	45.0	60.47	57.11	28.24	25.72
ChainScale_Independents	48.2	47.7	129.78	127.82	62.55	61.00
Class_Luxury	56.1	55.1	338.60	304.72	189.98	167.86
Class_Upper_Upscale	52.6	50.5	174.61	166.19	91.83	84.00
Class_Upscale	51.9	49.8	135.43	127.14	70.34	63.28
Class_Upper_Midscale	46.9	45.0	107.29	103.96	50.29	46.80
Class_Midscale	44.3	42.9	90.15	87.17	39.91	37.37
Class_Economy	46.9	45.6	69.45	67.01	32.54	30.58
Location_Urban	53.6	50.9	156.18	138.97	83.74	70.76
Location_Suburban	47.8	46.3	97.33	99.52	46.51	46.05
Location_Airport	58.4	56.2	104.45	100.02	60.98	56.25
Location_Interstate	41.3	40.6	78.41	75.22	32.38	30.55
Location_Resort	60.6	58.7	219.28	198.48	132.92	116.57
Location_Small_Metro/Town	39.7	38.3	95.21	89.70	37.81	34.34
Anaheim	70.6	64.7	150.20	134.26	105.97	86.85
Atlanta	55.3	53.7	105.77	94.97	58.53	51.04
Boston	48.1	41.5	137.49	128.05	66.12	53.17

Fin