General Salaries Overview

Code

CA Raul Marquez

26 noviembre 2022

Analyzing General Salaries Overview: US Market

With this report we seek to identify, for a human resources company, which positions have been most in demand in recent years, to identify the trend. What experience are they looking for, the residence of the employee, the possibility of working remotely.

📖 Background

All the information was obtained from the h1bdata.info portal, it was processed in PostgreSQL, and through a connection with R the analysis was carried out. 2000 positions were the basis of this analysis.

💾 The data

- Name of the position with vacancies.
- Number of vacancies available to fill.
- Average salaries of the positions.

The HR company is interested in knowing. - Where there are more vacancies, in which sectors. - What salaries do these vacancies have? - If there is a trend in any sector - General information that can help decision-making.

```
Error in install.packages : Updating loaded packages
```

```
Warning: package ‘rmarkdown’ was built under R version 4.2.1
```

```
Error in install.packages : Updating loaded packages
Error in install.packages : Updating loaded packages
Error in install.packages : Updating loaded packages
```

```
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate
version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into ‘C:/Users/rraul/AppData/Local/R/win-library/4.2’
(as ‘lib’ is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.2/sf_1.0-9.zip'
Content type 'application/zip' length 27951754 bytes (26.7 MB)
downloaded 26.7 MB
```

```
package ‘sf’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
    C:\Users\rraul\AppData\Local\Temp\Rtmp4q9eIL\downloaded_packages
```

```
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate
version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
```

```
Warning in install.packages :
  package ‘rmarkdown’ is in use and will not be installed
Error in install.packages : Updating loaded packages
Error in install.packages : Updating loaded packages
Error in install.packages : Updating loaded packages
```

```
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate
version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into ‘C:/Users/rraul/AppData/Local/R/win-library/4.2’
(as ‘lib’ is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.2/sqldf_0.4-11.zip'
Content type 'application/zip' length 78509 bytes (76 KB)
downloaded 76 KB
```

```
package ‘sqldf’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
    C:\Users\rraul\AppData\Local\Temp\Rtmp4q9eIL\downloaded_packages
```

```
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate
version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
```

```
Warning in install.packages :
  package ‘RPostgreSQL’ is in use and will not be installed
Error in install.packages : Updating loaded packages
Error in install.packages : Updating loaded packages
```

```
Warning: package ‘sf’ was built under R version 4.2.2Linking to GEOS 3.9.3, GDAL 3.5.2, PROJ 8.2.1; sf_use_s2() is TRUE
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate
version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
```

```
Warning in install.packages :
  package ‘RPostgres’ is in use and will not be installed
```

```
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate
version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
```

```
Warning in install.packages :
  package ‘DBI’ is in use and will not be installed
 [1] "exam"                           "data_salaries"
 [3] "salaries_data_analyst"          "salaries_all_positions"
 [5] "salaries_accountant"            "salaries_analyst"
 [7] "salaries_bi_analyst"            "salaries_business_analyst"
 [9] "salaries_data_scientist"        "salaries_financial_analyst"
[11] "salaries_financial_data_analyst" "salaries_staff_accountant"
```

	id	job_title	X._h1b_fill	avg_salary
	<S3: integer64>	<chr>	<chr>	<chr>
1	1	SOFTWARE ENGINEER	234,358	111,361
2	2	PROGRAMMER ANALYST	163,226	69,528
3	3	SOFTWARE DEVELOPER	149,795	88,065
4	4	SENIOR SOFTWARE ENGINEER	55,124	125,289
5	5	SYSTEMS ANALYST	43,107	74,656
6	6	SENIOR SYSTEMS ANALYST JC60	38,372	79,523

6 rows

```
Warning: package ‘tidyverse’ was built under R version 4.2.1Warning: package ‘ggplot2’ was built under R version 4.2.1Warnin
g: package ‘tibble’ was built under R version 4.2.1Warning: package ‘tidyr’ was built under R version 4.2.1Warning: package
‘readr’ was built under R version 4.2.1Warning: package ‘purrr’ was built under R version 4.2.1Warning: package ‘dplyr’ was
built under R version 4.2.1Warning: package ‘stringr’ was built under R version 4.2.1Warning: package ‘forcats’ was built un
der R version 4.2.1WARNING: Rtools is required to build R packages but is not currently installed. Please download and insta
ll the appropriate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.2/viridis_0.6.2.zip'
Content type 'application/zip' length 2999922 bytes (2.9 MB)
downloaded 2.9 MB
```

```
package ‘viridis’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
    C:\Users\rraul\AppData\Local\Temp\Rtmp4q9eIL\downloaded_packages
```

```
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate
version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.2/ggthemes_4.2.4.zip'
Content type 'application/zip' length 444308 bytes (433 KB)
downloaded 433 KB
```

```
package ‘ggthemes’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
    C:\Users\rraul\AppData\Local\Temp\Rtmp4q9eIL\downloaded_packages
```

```
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate
version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into ‘C:/Users/rraul/AppData/Local/R/win-library/4.2’
(as ‘lib’ is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.2/ggcharts_0.2.1.zip'
Content type 'application/zip' length 263275 bytes (257 KB)
downloaded 257 KB
```

```
package ‘ggcharts’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
    C:\Users\rraul\AppData\Local\Temp\Rtmp4q9eIL\downloaded_packages
```

```
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate
version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into ‘C:/Users/rraul/AppData/Local/R/win-library/4.2’
(as ‘lib’ is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.2/GGally_2.1.2.zip'
Content type 'application/zip' length 1636433 bytes (1.6 MB)
downloaded 1.6 MB
```

```
package ‘GGally’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
    C:\Users\rraul\AppData\Local\Temp\Rtmp4q9eIL\downloaded_packages
```

```
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate
version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into ‘C:/Users/rraul/AppData/Local/R/win-library/4.2’
(as ‘lib’ is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.2/ggcorrplot_0.1.4.zip'
Content type 'application/zip' length 31364 bytes (30 KB)
downloaded 30 KB
```

```
package ‘ggcorrplot’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
    C:\Users\rraul\AppData\Local\Temp\Rtmp4q9eIL\downloaded_packages
```

```
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate
version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into ‘C:/Users/rraul/AppData/Local/R/win-library/4.2’
(as ‘lib’ is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.2/formattable_0.2.1.zip'
Content type 'application/zip' length 173476 bytes (169 KB)
downloaded 169 KB
```

```
package ‘formattable’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
    C:\Users\rraul\AppData\Local\Temp\Rtmp4q9eIL\downloaded_packages
```

```
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate
version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into ‘C:/Users/rraul/AppData/Local/R/win-library/4.2’
(as ‘lib’ is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.2/flextable_0.8.3.zip'
Content type 'application/zip' length 1963559 bytes (1.9 MB)
downloaded 1.9 MB
```

```
package ‘flextable’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
    C:\Users\rraul\AppData\Local\Temp\Rtmp4q9eIL\downloaded_packages
```

```
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate
version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into ‘C:/Users/rraul/AppData/Local/R/win-library/4.2’
(as ‘lib’ is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.2/IRdisplay_1.1.zip'
Content type 'application/zip' length 34118 bytes (33 KB)
downloaded 33 KB
```

```
package ‘IRdisplay’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
    C:\Users\rraul\AppData\Local\Temp\Rtmp4q9eIL\downloaded_packages
```

```
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate
version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into ‘C:/Users/rraul/AppData/Local/R/win-library/4.2’
(as ‘lib’ is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.2/xtable_1.8-4.zip'
Content type 'application/zip' length 706980 bytes (690 KB)
downloaded 690 KB
```

```
package ‘xtable’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
    C:\Users\rraul\AppData\Local\Temp\Rtmp4q9eIL\downloaded_packages
```

```
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate
version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into ‘C:/Users/rraul/AppData/Local/R/win-library/4.2’
(as ‘lib’ is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.2/ggdark_0.2.1.zip'
Content type 'application/zip' length 1552170 bytes (1.5 MB)
downloaded 1.5 MB
```

```
package ‘ggdark’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
    C:\Users\rraul\AppData\Local\Temp\Rtmp4q9eIL\downloaded_packages
```

```
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate
version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into ‘C:/Users/rraul/AppData/Local/R/win-library/4.2’
(as ‘lib’ is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.2/mdthemes_0.1.0.zip'
Content type 'application/zip' length 59740 bytes (58 KB)
downloaded 58 KB
```

```
package ‘mdthemes’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
    C:\Users\rraul\AppData\Local\Temp\Rtmp4q9eIL\downloaded_packages
```

```
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate
version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into ‘C:/Users/rraul/AppData/Local/R/win-library/4.2’
(as ‘lib’ is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.2/ggExtra_0.10.0.zip'
Content type 'application/zip' length 363465 bytes (354 KB)
downloaded 354 KB
```

```
package ‘ggExtra’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
    C:\Users\rraul\AppData\Local\Temp\Rtmp4q9eIL\downloaded_packages
```

```
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate
version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into ‘C:/Users/rraul/AppData/Local/R/win-library/4.2’
(as ‘lib’ is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.2/ggtext_0.1.2.zip'
Content type 'application/zip' length 1259317 bytes (1.2 MB)
downloaded 1.2 MB
```

```
package ‘ggtext’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
    C:\Users\rraul\AppData\Local\Temp\Rtmp4q9eIL\downloaded_packages
```

```
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate
version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into ‘C:/Users/rraul/AppData/Local/R/win-library/4.2’
(as ‘lib’ is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.2/waffle_0.7.0.zip'
Content type 'application/zip' length 313241 bytes (305 KB)
downloaded 305 KB
```

```
package ‘waffle’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
    C:\Users\rraul\AppData\Local\Temp\Rtmp4q9eIL\downloaded_packages
```

```
Warning: package ‘viridis’ was built under R version 4.2.2Warning: package ‘viridisLite’ was built under R version 4.2.1Warn
ing: package ‘ggthemes’ was built under R version 4.2.2Warning: package ‘GGally’ was built under R version 4.2.2Warning: pac
kage ‘ggcharts’ was built under R version 4.2.2Warning: package ‘ggcorrplot’ was built under R version 4.2.2Warning: package
‘patchwork’ was built under R version 4.2.1Warning: package ‘formattable’ was built under R version 4.2.2Warning: package ‘f
lextable’ was built under R version 4.2.2Warning: package ‘IRdisplay’ was built under R version 4.2.2Warning: package ‘xtabl
e’ was built under R version 4.2.2Warning: package ‘ggdark’ was built under R version 4.2.2Warning: package ‘mdthemes’ was b
uilt under R version 4.2.2Warning: package ‘ggExtra’ was built under R version 4.2.2Warning: package ‘ggtext’ was built unde
r R version 4.2.2Warning: package ‘waffle’ was built under R version 4.2.2
```

Comencemos por crear categorias, voy a tener que crear asi e ir repitiendo p/ luego pasarlo a el dataframe

unimos en una columna todos los valores encontrados

	job_title <chr>	id <S3: integer64>	cat <chr>
1	software engineer	1	engineer
2	programmer analyst	2	engineer
3	software developer	3	engineer
4	senior software engineer	4	engineer

	job_title <chr>	id <S3: integer64>	cat <chr>
5	systems analyst	5	engineer
6	senior systems analyst jc60	6	engineer

6 rows

```
'data.frame':   2401 obs. of  3 variables:
 $ job_title: chr  "software engineer" "programmer analyst" "software developer" "senior software engineer" ...
 $ id       :integer64 1 2 3 4 5 6 7 9 ...
 $ cat      : chr  "engineer" "engineer" "engineer" "engineer" ...
```

Vamos a preparar la data, renombrar y q sean numeros avg_salary y X._H1B_Fill

	id <S3: integer64>	job_title <chr>	X._h1b_fill <chr>	avg_salary <chr>
1	1	SOFTWARE ENGINEER	234,358	111,361
2	2	PROGRAMMER ANALYST	163,226	69,528
3	3	SOFTWARE DEVELOPER	149,795	88,065
4	4	SENIOR SOFTWARE ENGINEER	55,124	125,289
5	5	SYSTEMS ANALYST	43,107	74,656
6	6	SENIOR SYSTEMS ANALYST JC60	38,372	79,523

6 rows

```
'data.frame':	2000 obs. of  4 variables:
 $ id                :integer64 1 2 3 4 5 6 7 8 ...
 $ job_title         : chr  "SOFTWARE ENGINEER
" "PROGRAMMER ANALYST                                                                                " "SOFTWARE DEVELOPER
" "SENIOR SOFTWARE ENGINEER                                                                          " ...
 $ positions_available: chr  "234,358
" "163,226                                                                                           " "149,795
" "55,124                                                                                            " ...
 $ avg_salary        : chr  "111,361
" "69,528                                                                                            " "88,065
" "125,289                                                                                           " ...
```

	id <S3: integer64>	job_title <chr>	positions_available <dbl>	avg_salary <dbl>
1	1	software engineer	234358	111361
2	2	programmer analyst	163226	69528
3	3	software developer	149795	88065
4	4	senior software engineer	55124	125289
5	5	systems analyst	43107	74656
6	6	senior systems analyst jc60	38372	79523

6 rows

id <S3: integer64>	job_title <chr>	positions_available <dbl>
184	computer programmer configurer 2	2469
295	computer specialist testing and quality analyst 2	1499
451	computer programmer configurer 3	924
683	computer specialist testing and quality analyst 3	573
721	programmer analyst	537

id	job_title	positions_available
<S3: integer64>	<chr>	<dbl>
789	computer programmer analyst	477
870	computer programmer configurer 1	424
1254	computer specialist system support and development admin 2	258
1560	computer specialist testing and quality analyst 1	199

9 rows | 1-3 of 4 columns

no encontramos duplicados en el proceso

armamos algunos categorias grupales, que abarquen varios trabajos distintos

```
'data.frame':	2534 obs. of  6 variables:
 $ id                 :integer64 1 2 2 3 4 5 5 6 ...
 $ job_title.x        : chr  "software engineer
" "programmer analyst                                                                                    " "programmer analyst
" "software developer                                                                                    " ...
 $ positions_available: num  234358 163226 163226 149795 55124 ...
 $ avg_salary         : num  111361 69528 69528 88065 125289 ...
 $ job_title.y        : chr  "software engineer" "programmer analyst" "programmer analyst" "software developer" ...
 $ cat                : chr  "engineer" "engineer" "data" "engineer" ...
```

n
<int>
133

1 row

	id	job_title.x	positions_available	avg_salary	job_title.y
	<S3: integer64>	<chr>	<dbl>	<dbl>	<chr>
1	1	software engineer	234358	111361	software engineer
2	2	programmer analyst	163226	69528	programmer analyst

	id	job_title.x	positions_available	avg_salary	job_title.y
	<S3: integer64>	<chr>	<dbl>	<dbl>	<chr>
3	2	programmer analyst	163226	69528	programmer analyst
4	3	software developer	149795	88065	software developer
5	4	senior software engineer	55124	125289	senior software engineer
6	5	systems analyst	43107	74656	systems analyst

6 rows | 1-6 of 6 columns

```
'data.frame':	2534 obs. of  6 variables:
 $ id                 :integer64 1 2 2 3 4 5 5 6 ...
 $ job_title.x        : chr  "software engineer
" "programmer analyst                                                                        " "programmer analyst
" "software developer                                                                        " ...
 $ positions_available: num  234358 163226 163226 149795 55124 ...
 $ avg_salary         : num  111361 69528 69528 88065 125289 ...
 $ job_title.y        : chr  "software engineer" "programmer analyst" "programmer analyst" "software developer" ...
 $ cat                : chr  "engineer" "engineer" "data" "engineer" ...
[1] 543
```

id	job_title.x	positions_available
<S3: integer64>	<chr>	<dbl>
2	programmer analyst	163226
5	systems analyst	43107
6	senior systems analyst jc60	38372
8	business analyst	33275
13	computer systems analyst	26117
17	project manager	22862
25	technology analyst - us	14579
32	computer systems analysts	10801

id <S3: integer64>	job_title.x <chr>	positions_available <dbl>
34	business systems analyst	10273
34	business systems analyst	10273

1-10 of 675 rows | 1-3 of 6 columns Previous **1** 2 3 4 5 6 … 68 Next

```
'data.frame':   1991 obs. of  6 variables:
 $ id                :integer64 1 2 3 4 5 6 7 8 ...
 $ job_title.x       : chr  "software engineer
" "programmer analyst                                                                  " "software developer
" "senior software engineer                                                            " ...
 $ positions_available: num  234358 163226 149795 55124 43107 ...
 $ avg_salary        : num  111361 69528 88065 125289 74656 ...
 $ job_title.y       : chr  "software engineer" "programmer analyst" "software developer" "senior software engineer" ...
 $ cat               : chr  "engineer" "engineer" "engineer" "engineer" ...
[1] 133
```

limpiamos los NA en cat para que sean other

```
% latex table generated in R 4.2.0 by xtable 1.8-4 package
% Sat Nov 26 11:14:41 2022
\begin{table}[ht]
\centering
\begin{tabular}{rllllll}
  \hline
 &       id & job\_title.x & positions\_available &   avg\_salary &     cat \\
  \hline
X & Min.   :   1   & Length:1991        & Min.   :     3.0   & Min.   :    78   & Length:1991        \\
  X.1 & 1st Qu.: 501   & Class :character   & 1st Qu.:   205.0   & 1st Qu.: 72351   & Class :character   \\
  X.2 & Median :1003   & Mode  :character   & Median :   329.0   & Median : 89593   & Mode  :character   \\
  X.3 & Mean   :1001   &  & Mean   :  1396.0   & Mean   : 90952   &  \\
  X.4 & 3rd Qu.:1502   &  & 3rd Qu.:   755.5   & 3rd Qu.:111506   &  \\
  X.5 & Max.   :2000   &  & Max.   :234358.0   & Max.   :434946   &  \\
   \hline
\end{tabular}
\end{table}
```

cat <chr>	mean_salary <dbl>	mean_positions <dbl>
data	81633.36	947.3309
engineer	90430.97	1592.7747
finance	86528.54	1099.2347
medicine	145077.00	706.0000
other	97120.63	1080.0872
university	78106.20	1086.8522

6 rows

sacamos los valores pequenos irracionales

saquemos los other que son medicine

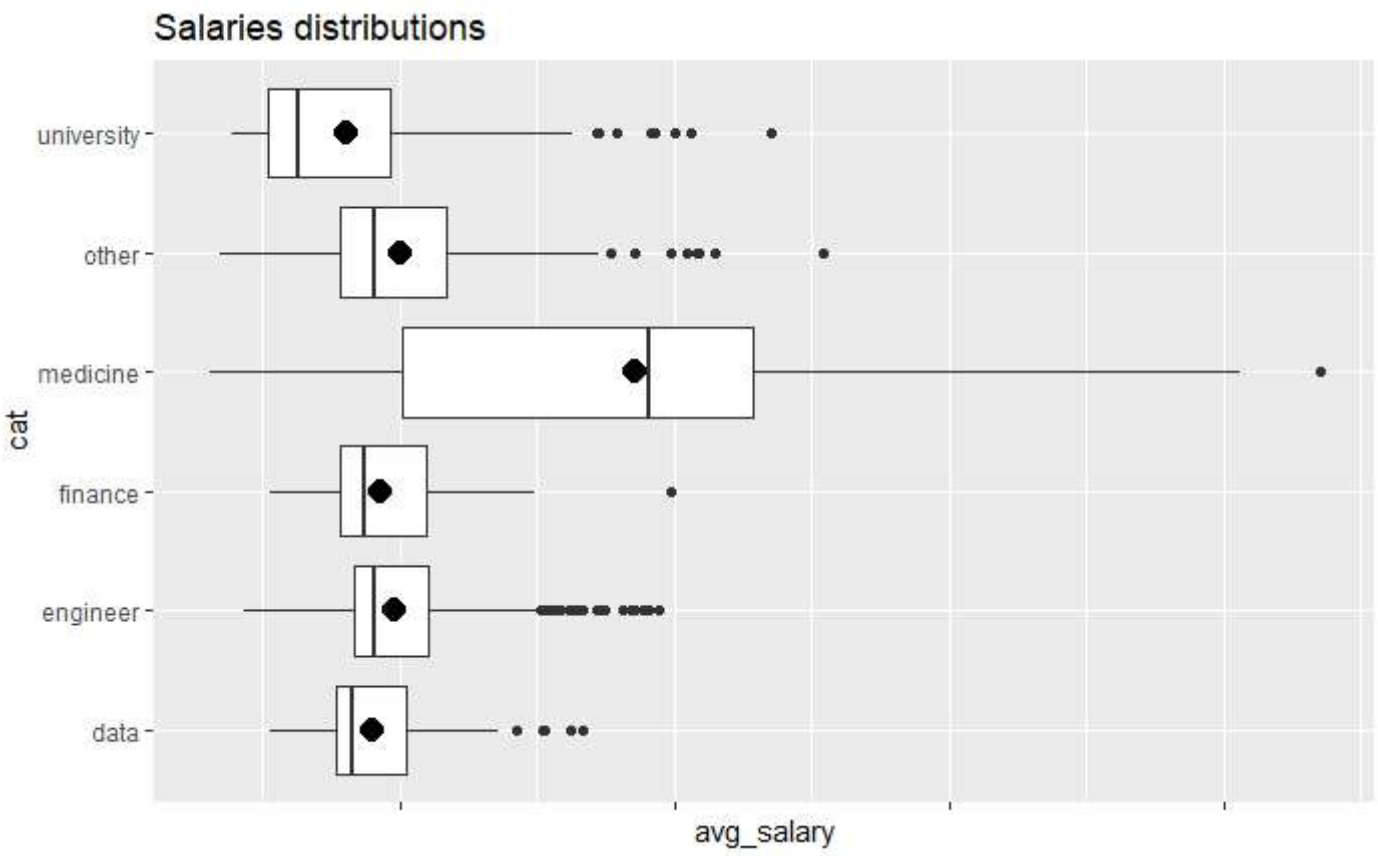
id <S3: integer64>	job_title.x <chr>	positions_available <dbl>	avg_salary <dbl>
114	market research analyst	4295	56669
751	market analyst	514	56922
1207	marketing research analyst	268	57405
483	solution analyst	859	58546
552	budget analyst	727	62166
1986	progammer analyst	148	62476
641	logistics analyst	617	63363
1290	accounting analyst	248	63563
399	marketing analyst	1066	63566
1973	delivery analyst 2	149	63927

1-10 of 97 rows | 1-4 of 5 columns

cat <chr>	mean(avg_salary) <dbl>	median <dbl>	IQR(avg_salary) <dbl>	quantile(avg_salary, 0.25) <dbl>
medicine	185950.60	190413.0	127475.50	101067.25
engineer	97959.43	91004.0	27062.00	83394.50
other	99971.93	90922.0	38835.50	78299.50
finance	92853.96	86926.5	31028.25	78332.25
data	90203.12	82758.0	25589.50	76841.50
university	80399.61	63001.0	44497.50	52132.00

6 rows

si observamos los datos, entendemos que la media de medicina de $185951 es una excelente herramienta para interpetar los datos, pero en el caso de las otras categorias que se encuentran sesgadas a la derecha, otras medidas, la mediana nos puede brindar informacion un poco mas valiosa, sobre todo sin consideramos el IQR. Pero claramente una oferta para las posiciones indicadas deben estar alrededor de la media para ser interesantes para candidatos destacados.




```
      id       job_title.x        positions_available   avg_salary
Min.   :  10   Length:115         Min.   :  121.0     Min.   : 38499
1st Qu.: 561   Class :character   1st Qu.:  222.0     1st Qu.: 52132
Median : 986   Mode  :character   Median :  344.0     Median : 63001
Mean   : 993                      Mean   : 1086.8     Mean   : 80400
3rd Qu.:1405                      3rd Qu.:  709.5     3rd Qu.: 96630
Max.   :2000                      Max.   :31798.0     Max.   :235027
    cat
Length:115
Class :character
Mode  :character
```

cat <chr>	suma <dbl>	percent <dbl>
data	131679	4
engineer	2037343	73
finance	107725	3
medicine	36793	1
other	340812	12
university	124988	4

6 rows

cat <chr>	mean <dbl>	median <dbl>	iqr <dbl>	quantile <dbl>	max <dbl>	sd <dbl>
university	1086.8522	344.0	487.50	222.00	31798	3228.4102
medicine	707.5577	334.5	497.50	219.00	5028	917.9621
engineer	1592.9187	332.0	562.00	207.00	234358	9577.9313
finance	1099.2347	322.0	470.25	218.25	33275	3519.0789
other	1106.5325	311.0	575.00	193.75	27533	3065.4343
data	947.3309	295.0	545.00	187.00	23681	2541.8925

6 rows

la tabla 2 nos indica las posiciones disponibles de cada posicion por categoria en promedio tambien al haber posiciones extremas, la mediana vuelve a resultar una mejor medida para entender cuantas posiciones hay disponibles para cada posicion, un analisis mas profundo de como se desarrollan sin esos valores extremos podria ayudarnos a entender mas. aqui nuevamente la curva de posiciones disponibles se encuentra sesgada a la derecha con una gran concentracion cercana a la mediana.

	id <S3: integer64>	job_title.x <chr>	positions_available <dbl>	avg_salary <dbl>	cat <chr>
1	6	senior systems analyst jc60	38372	79523	engineer

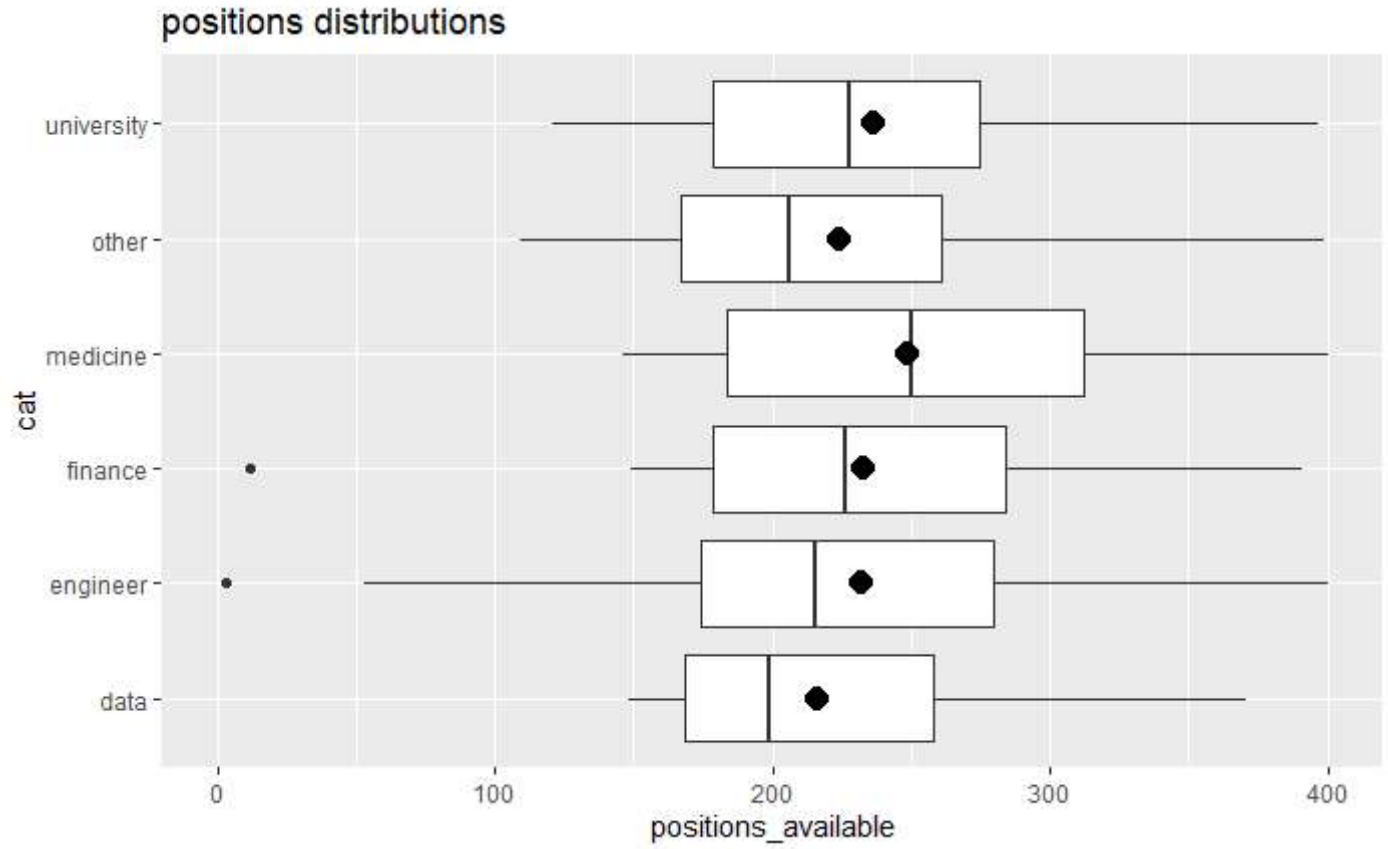
	id	job_title.x	positions_available	avg_salary	cat
	<S3: integer64>	<chr>	<dbl>	<dbl>	<chr>
2	7	computer programmer	34945	90431	engineer
3	8	business analyst	33275	74979	finance
4	9	developer	32082	90431	engineer
5	10	assistant professor	31798	109364	university
6	11	manager jc50	27533	96457	other
7	12	consultant	26334	82925	other
8	13	computer systems analyst	26117	75187	engineer
9	14	senior consultant	24341	95692	other
10	15	technology lead - us	23959	80803	engineer

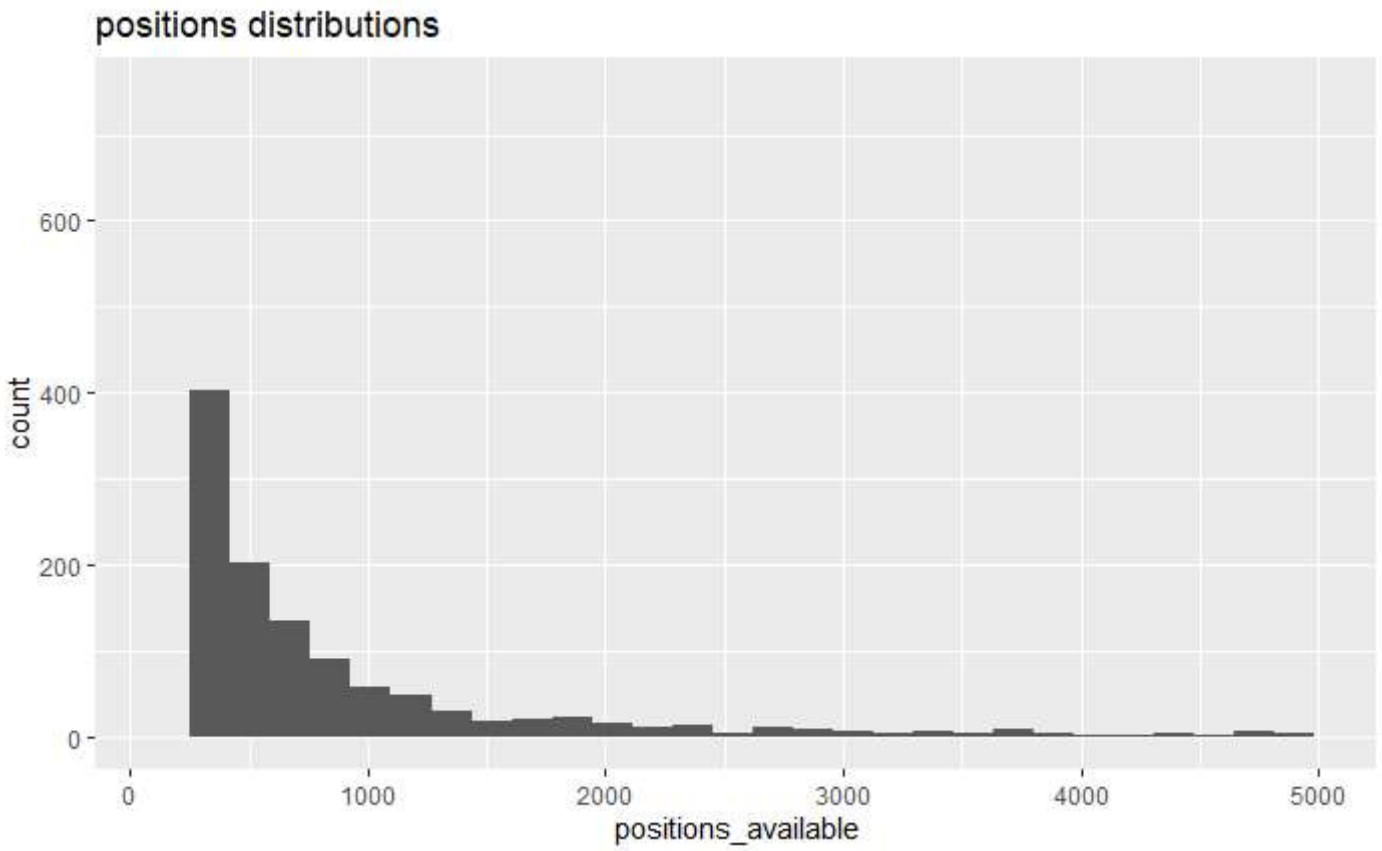
1-10 of 20 rows Previous **1** 2 Next

	id	job_title.x	positions_available	avg_salary	cat
	<S3: integer64>	<chr>	<dbl>	<dbl>	<chr>
1	1	software engineer	234358	111361	engineer
2	2	programmer analyst	163226	69528	engineer
3	3	software developer	149795	88065	engineer
4	4	senior software engineer	55124	125289	engineer
5	5	systems analyst	43107	74656	engineer
6	6	senior systems analyst jc60	38372	79523	engineer
7	7	computer programmer	34945	90431	engineer
8	8	business analyst	33275	74979	finance
9	9	developer	32082	90431	engineer
10	10	assistant professor	31798	109364	university

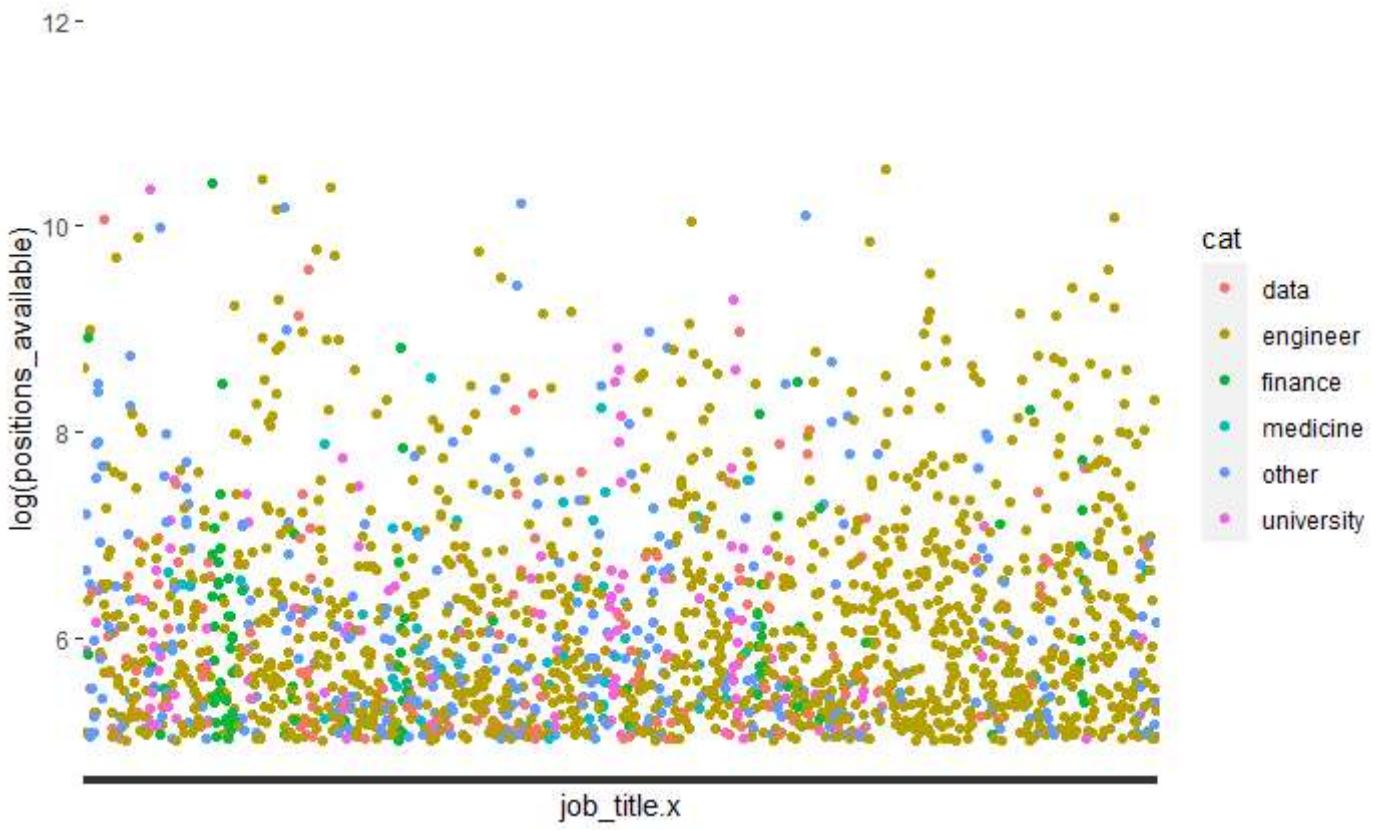
existe una clara diferencia en la cantidad de puestos disponibles para posiciones de IT Se destacan las posiciones de sofware engineer, programmer analyst, software developer, recien en octava posicion encontramos otra rama la de business analyst, siendo igualmente el 14% de la cantidad de posiciones demandadas de software engineer.

```
Scale for 'y' is already present. Adding another scale for 'y', which will replace the
existing scale.
```





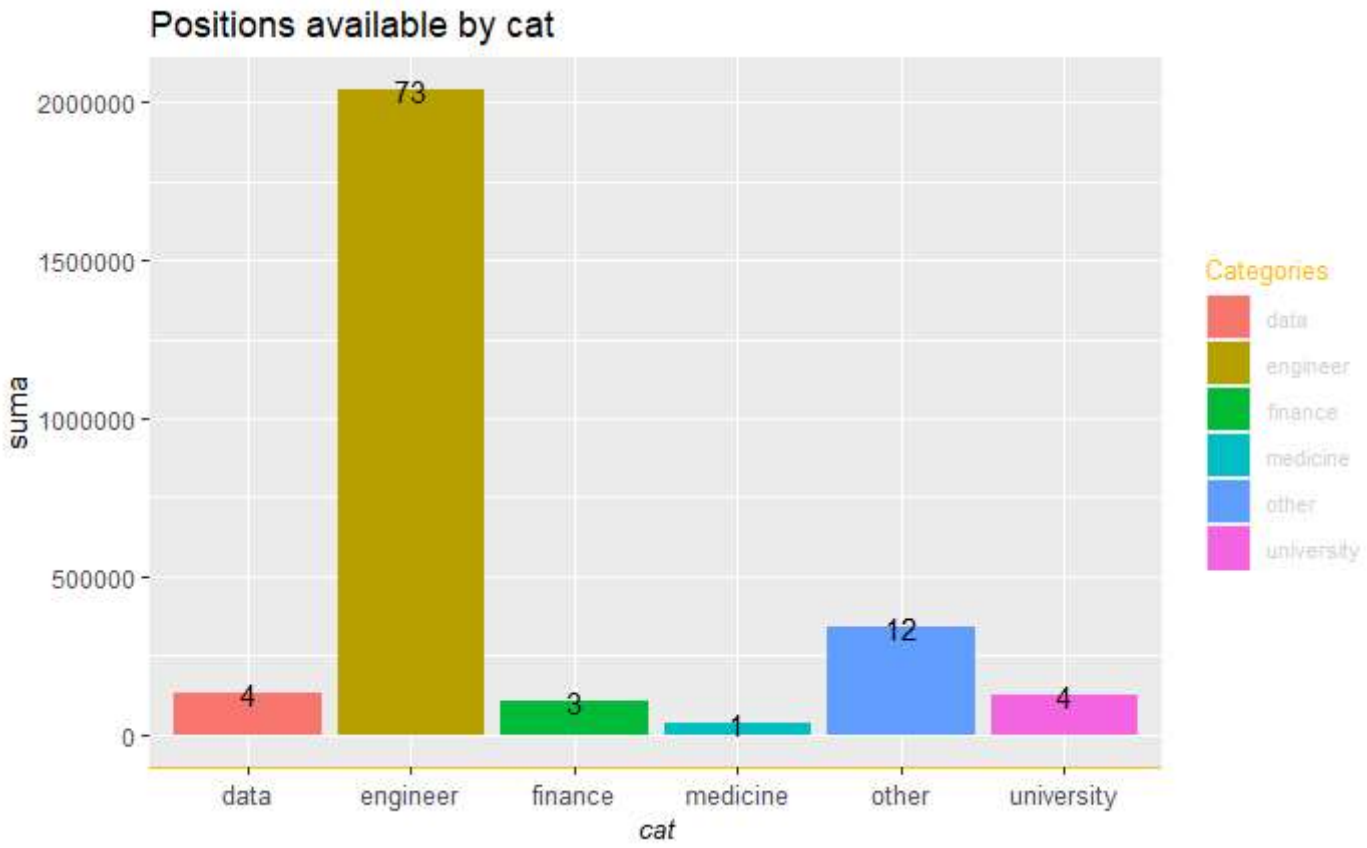

```
Scale for 'y' is already present. Adding another scale for 'y', which will replace the
existing scale.
```



quantile
<dbl>
9185.4

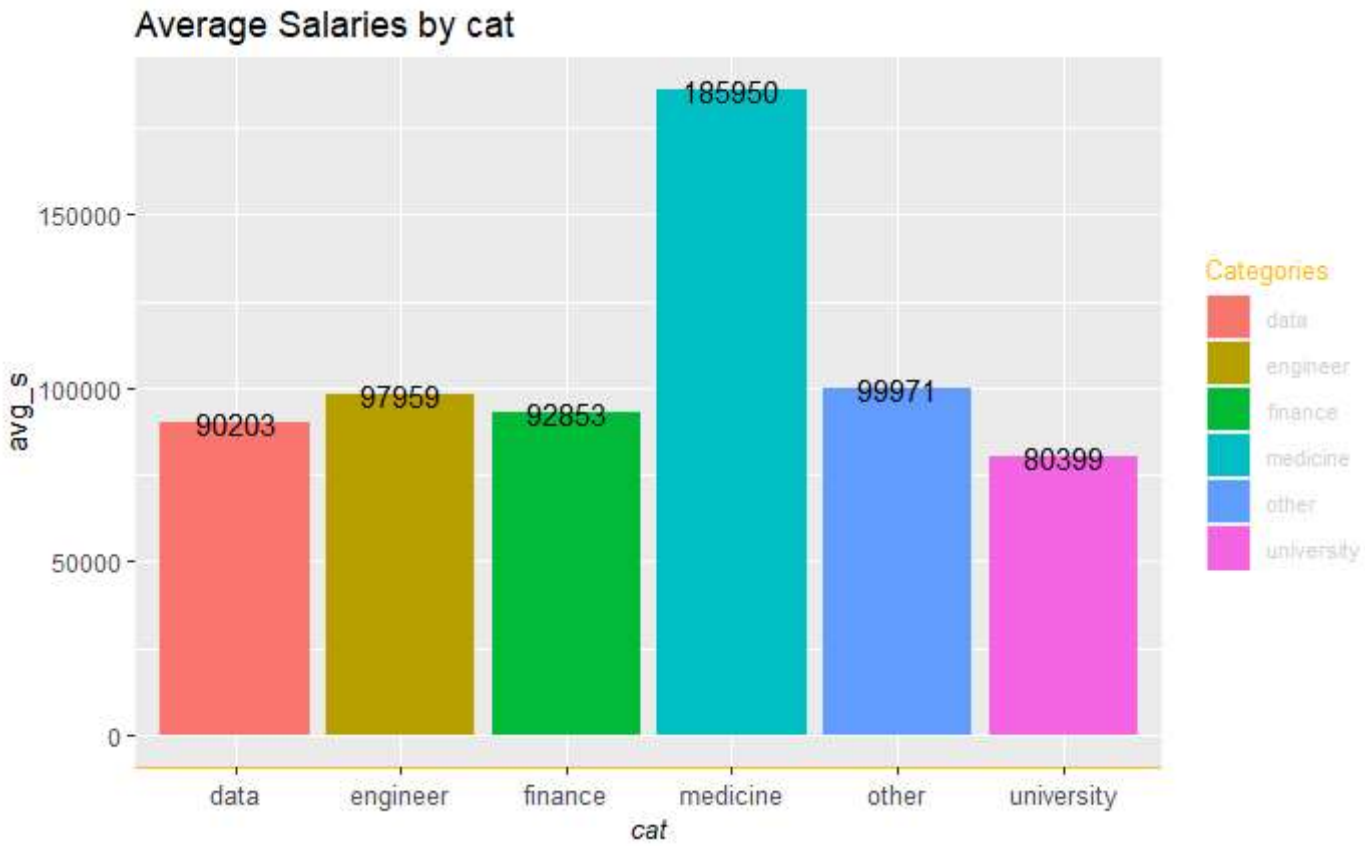
1 row

La gran dispersion de la cantidad de posiciones, no permite observar con los valores de sd y mean, claramente como se distribuyen. El 80% de las posiciones, tienen menos de 952 puestos disponibles, solo el 10% de posiciones mas demandadas tienen mas de 2000 lugares y solamente el 2% de las posiciones mas demandadas tienen mas de 9000 disponibles.



cat <chr>	avg_s <dbl>	percent <dbl>
data	90203	13
engineer	97959	15
finance	92853	14
medicine	185950	28
other	99971	15
university	80399	12

6 rows



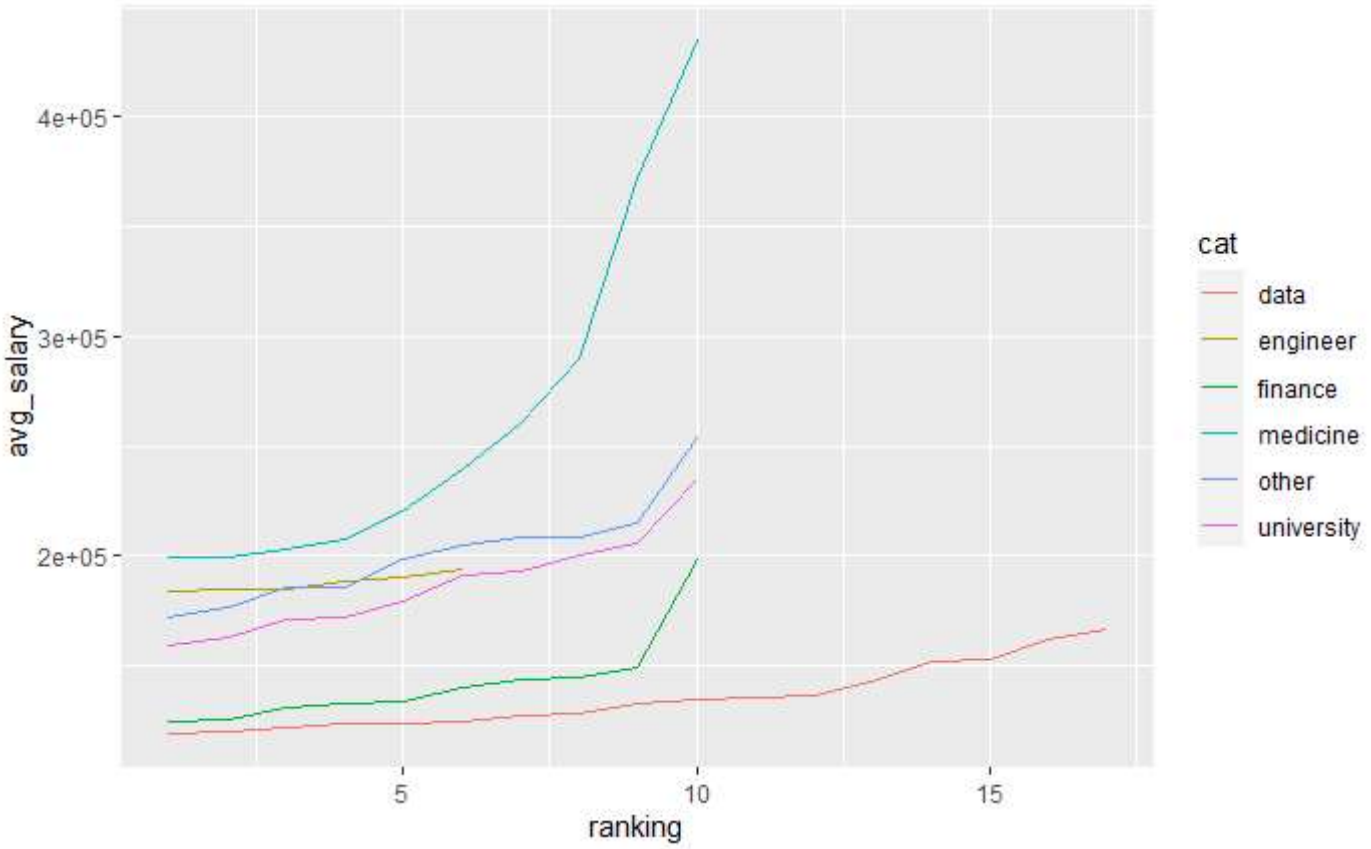
We can see that, salaries have an Average of $103000, and that most average salaries are around that, except for medicine sector that has the most significant average salaries.

mean(avg_s)
<dbl>
107889.2

1 row

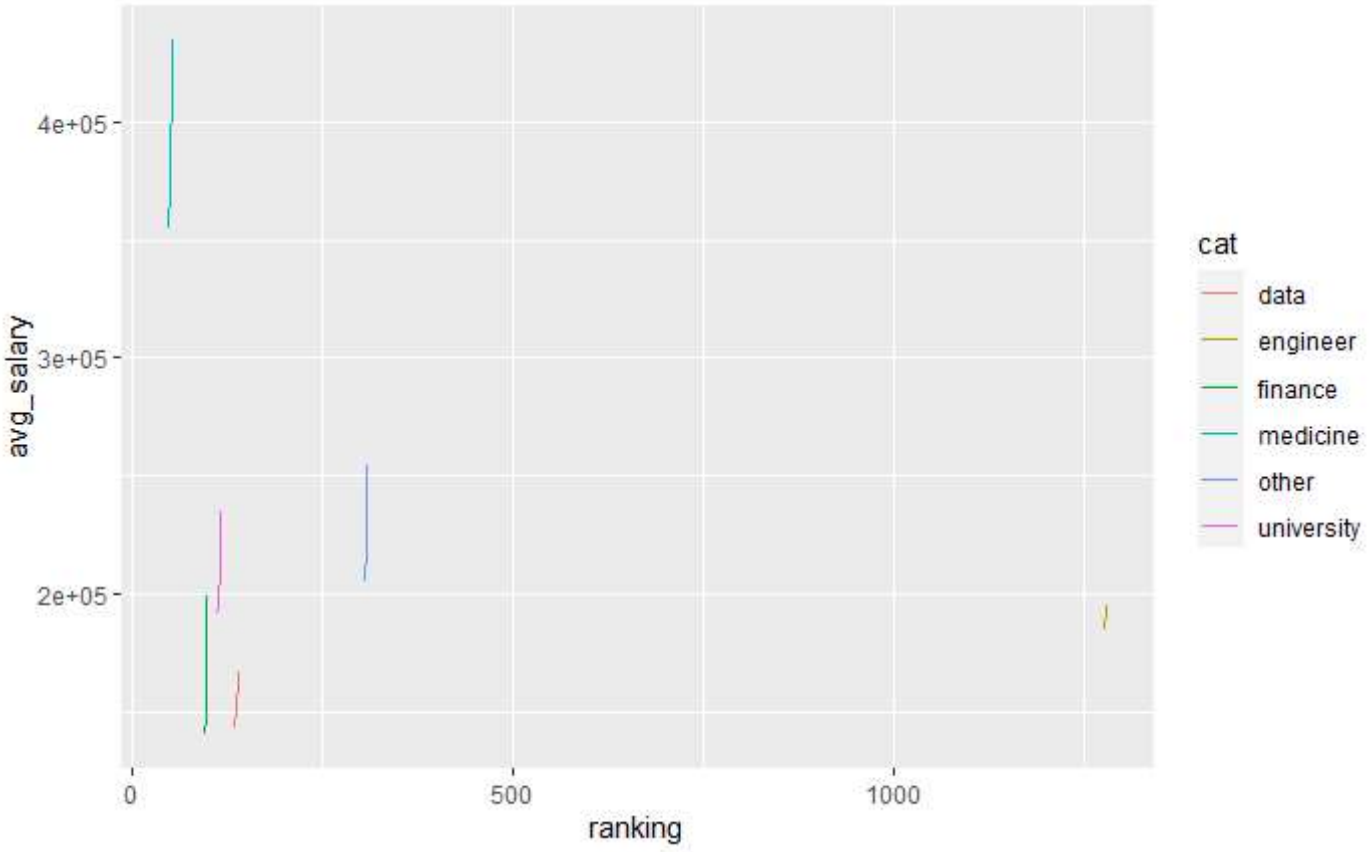
job_title.x	positions_available	cat	avg_salary	ranking
<chr>	<dbl>	<chr>	<dbl>	<dbl>
data science manager	295	data	166346	17
applied scientist iii	193	data	162453	16

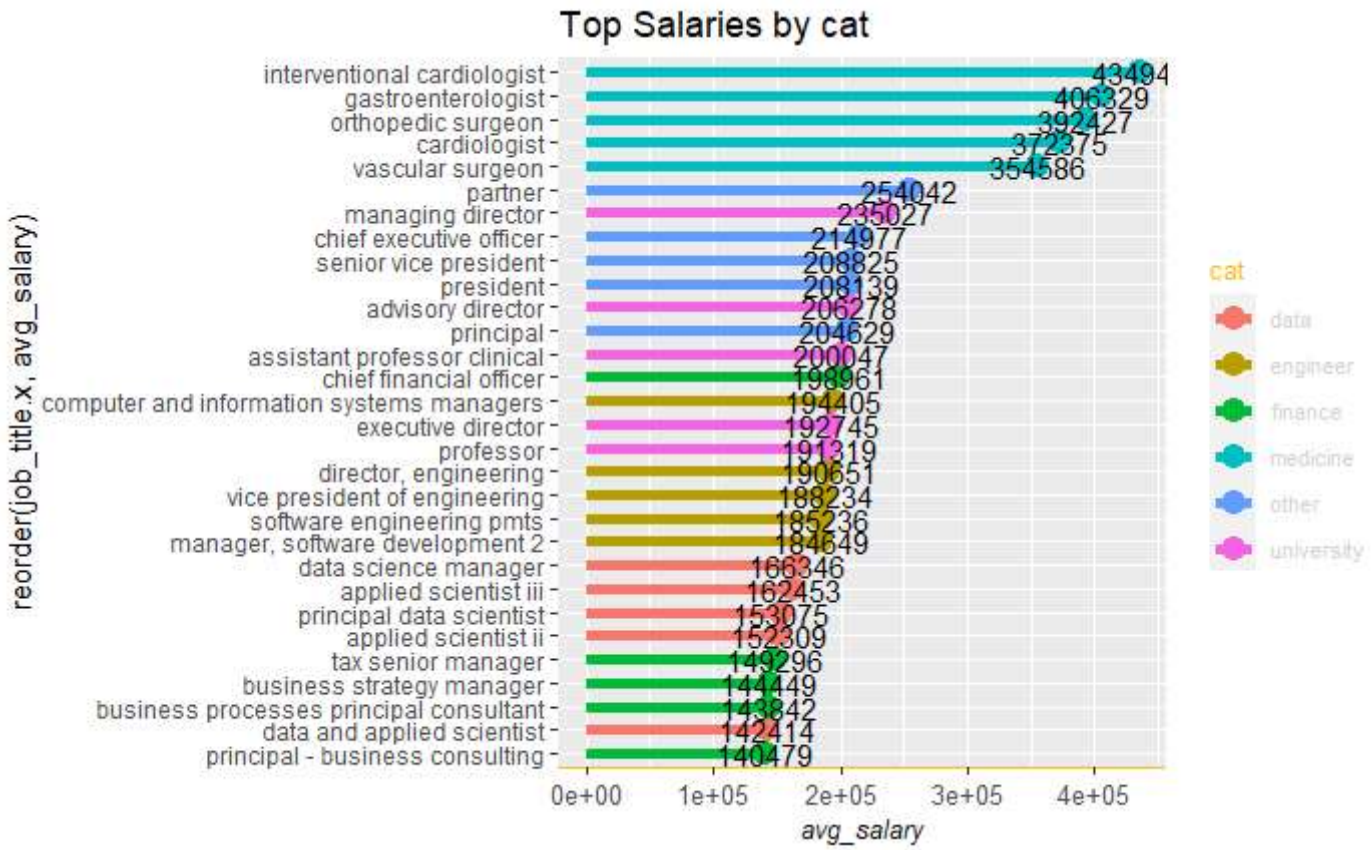
job_title.x <chr>	positions_available <dbl>	cat <chr>	avg_salary <dbl>	ranking <dbl>
principal data scientist	259	data	153075	15
applied scientist ii	1014	data	152309	14
data and applied scientist	1615	data	142414	13
applied scientist i	595	data	136235	12
senior data scientist	3034	data	135579	11
applied scientist	293	data	134197	10
lead data scientist	418	data	132654	9
machine learning scientist	162	data	128152	8



job_title.x <chr>	positions_available <dbl>	cat <chr>	avg_salary <dbl>	ranking <dbl>
data science manager	295	data	166346	17
applied scientist iii	193	data	162453	16
principal data scientist	259	data	153075	15
applied scientist ii	1014	data	152309	14
data and applied scientist	1615	data	142414	13
applied scientist i	595	data	136235	12
senior data scientist	3034	data	135579	11

job_title.x <chr>	positions_available <dbl>	cat <chr>	avg_salary <dbl>	ranking <dbl>
applied scientist	293	data	134197	10
lead data scientist	418	data	132654	9
machine learning scientist	162	data	128152	8

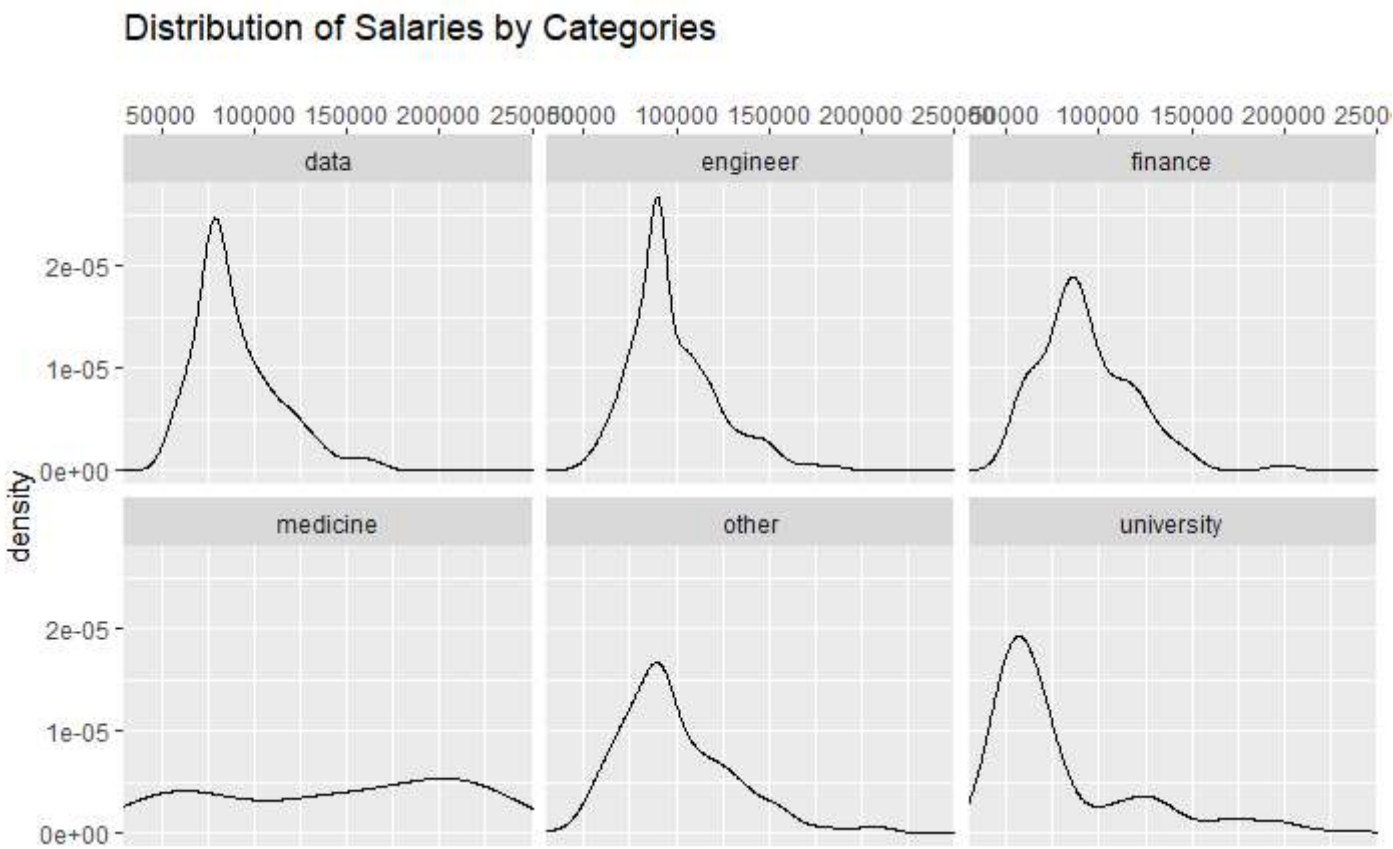




job_title.x
<chr>
computer and information systems managers
director, engineering
vice president of engineering
software engineering pmts
manager, software development 2

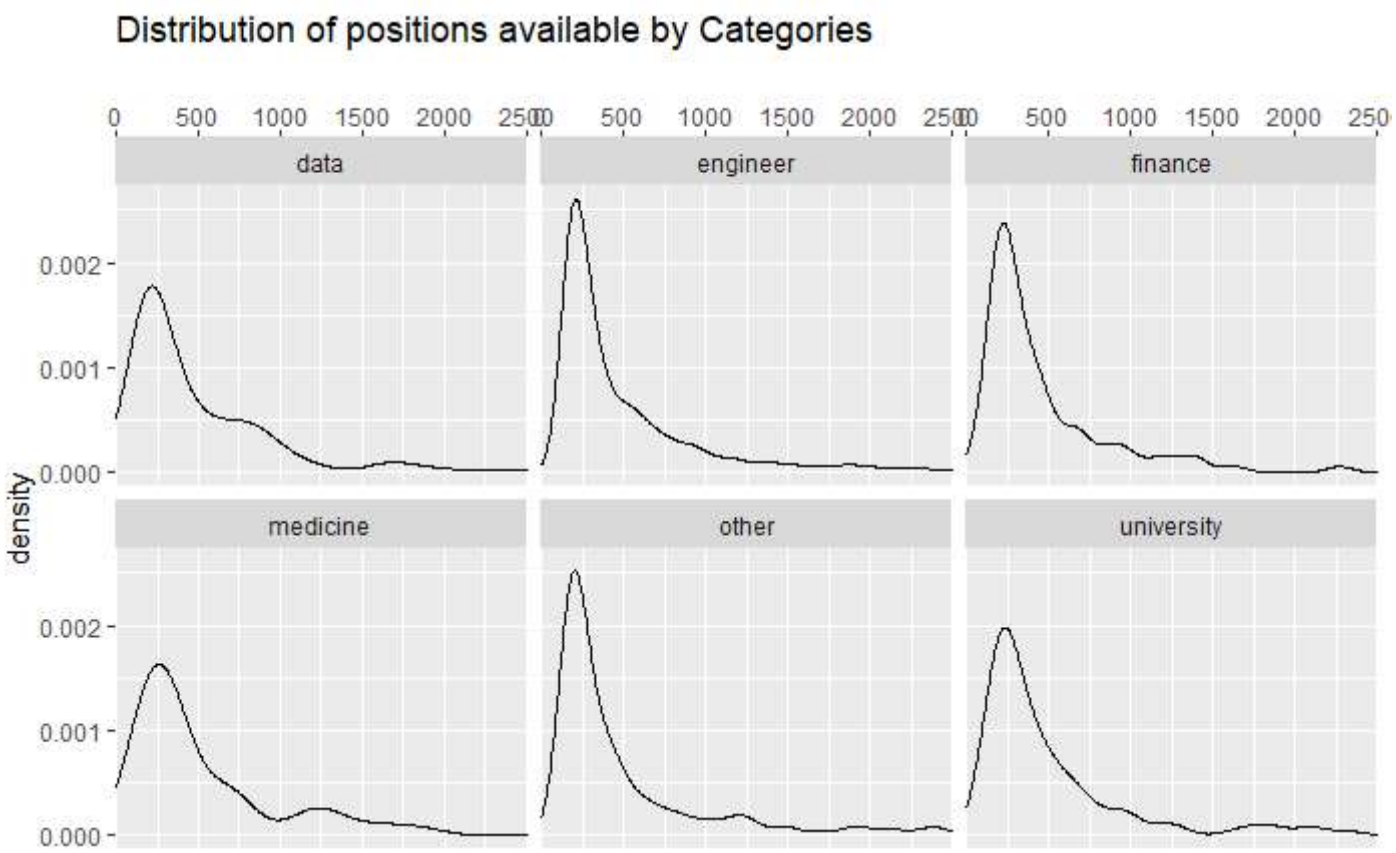
5 rows

Salaries density by categories.



salaries have a great concentration in 100 k or less, with only medicine sector receiving less concentration and with their professionals with less density in lower incomes

positions available density

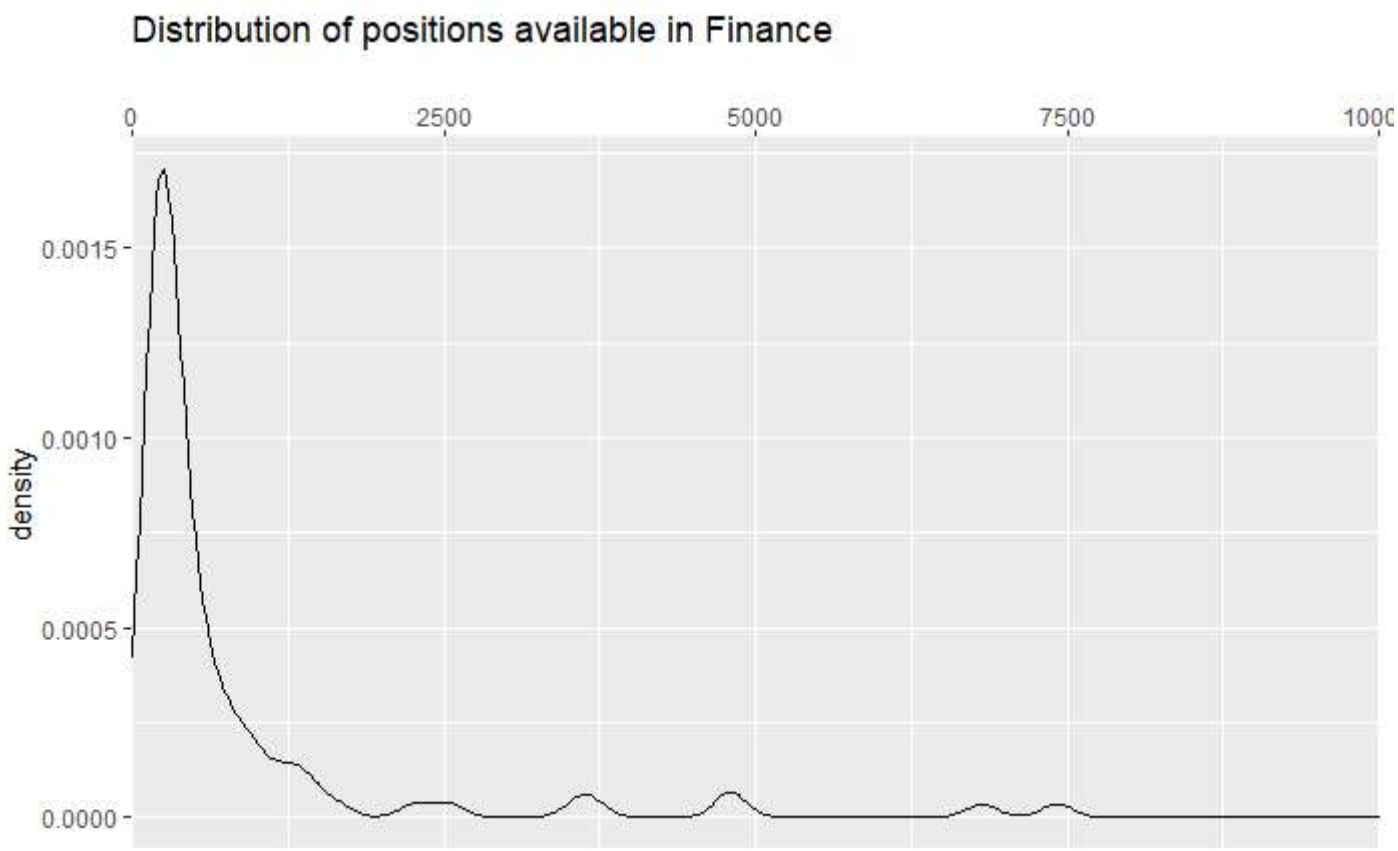



```
'data.frame':   1991 obs. of  5 variables:
 $ id                 :integer64 1 2 3 4 5 6 7 8 ...
 $ job_title.x        : chr  "software engineer" "programmer analyst" "software developer" "senior software engineer" ...
 $ positions_available: num  234358 163226 149795 55124 43107 ...
 $ avg_salary         : num  111361 69528 88065 125289 74656 ...
 $ cat                : chr  "engineer" "engineer" "engineer" "engineer" ...
```

como se comportan las posiciones y salarios de finanzas y datos sacar posiciones demandadas dentro de finance y data las mas pedidas.

cat	mean	median	iqr	quantile	max	sd
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
finance	1099.235	322	470.25	218.25	33275	3519.079

1 row



cat	quantile(positions_available, 0.8)
<chr>	<dbl>
finance	879.2

1 row

En finance hay 98 posiciones el 90 % de las posiciones tienen menos de 1500 puestos disponibles y el 80 % menos de 880 posiciones, este datos es fundamental para definir la oferta disponible de puestos.Solo Business Analyst, una vacante que se cruza con data e Engineer posee mas de 33 mil posiciones disponibles.

id	job_title.x	positions_available	avg_salary	cat
<S3: integer64>	<chr>	<dbl>	<dbl>	<chr>
8	business analyst	33275	74979	finance

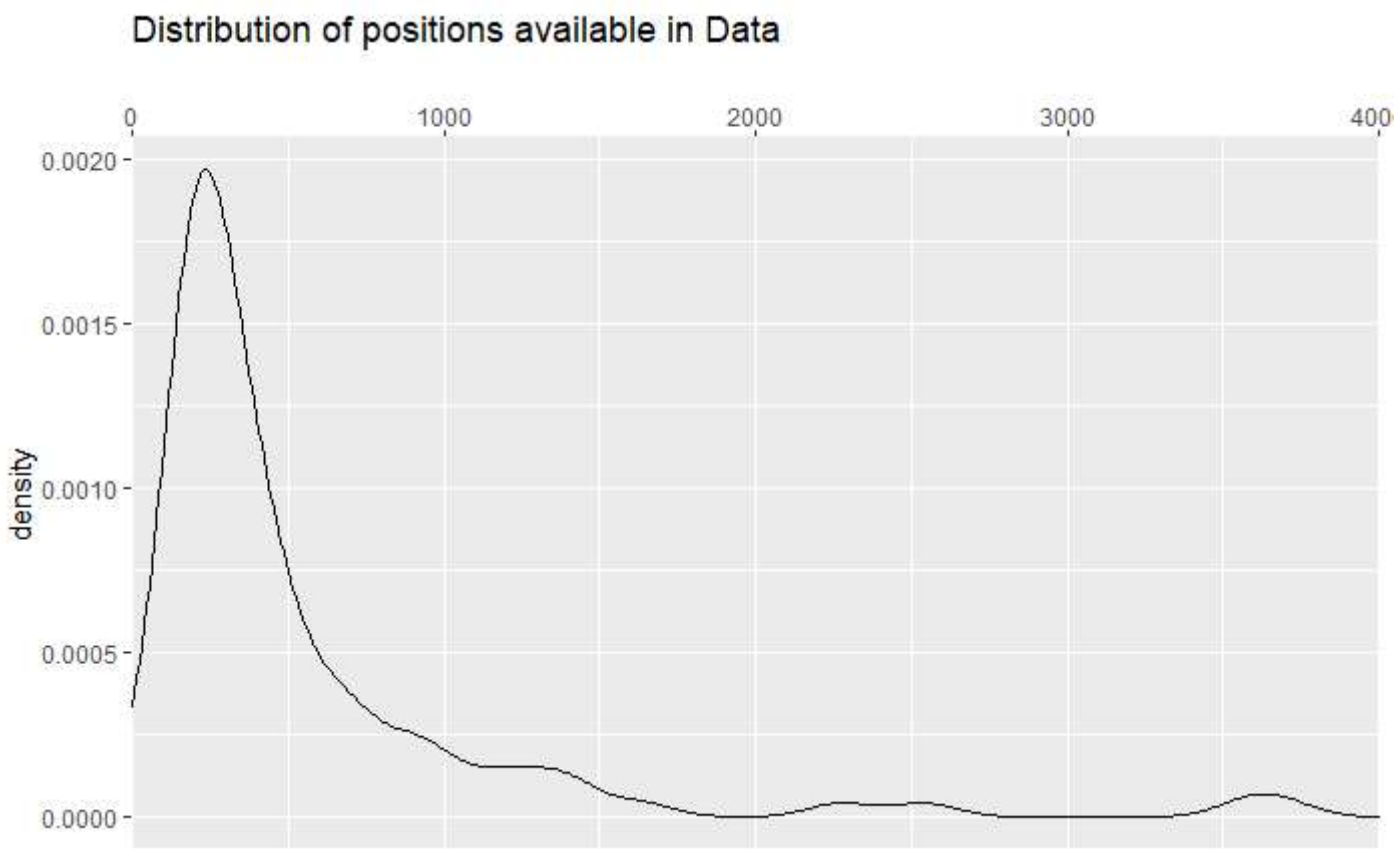
id	job_title.x	positions_available	avg_salary	cat
<S3: integer64>	<chr>	<dbl>	<dbl>	<chr>
53	accountant	7416	59416	finance
59	financial analyst	6819	79152	finance
100	senior business analyst	4809	97067	finance
102	business intelligence analyst	4792	80011	finance
127	staff accountant	3693	56768	finance
132	sap consultant	3573	86529	finance
181	financial analysts	2543	86529	finance
204	tax senior	2271	78278	finance
272	business development manager	1632	101283	finance

1-10 of 20 rows Previous **1** 2 Next

```
grouped_df [98 × 5] (S3: grouped_df/tbl_df/tbl/data.frame)
 $ id                :integer64 [1:98] 8 53 59 100 102 127 132 181 ...
 $ job_title.x       : chr [1:98] "business analyst" "accountant" "financial analyst" "senior business analyst" ...
 $ positions_available: num [1:98] 33275 7416 6819 4809 4792 ...
 $ avg_salary        : num [1:98] 74979 59416 79152 97067 80011 ...
 $ cat               : chr [1:98] "finance" "finance" "finance" "finance" ...
 - attr(*, "groups")= tibble [1 × 2] (S3: tbl_df/tbl/data.frame)
  ..$ cat  : chr "finance"
  ..$ .rows: list<int> [1:1]
  .. ..$ : int [1:98] 1 2 3 4 5 6 7 8 9 10 ...
  .. ..@ ptype: int(0)
  ..- attr(*, ".drop")= logi TRUE
```

cat	mean	median	iqr	quantile	max	sd
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
data	947.3309	295	545	187	23681	2541.892

1 row



cat	quantile(positions_available, 0.99)
<chr>	<dbl>
data	12391.2

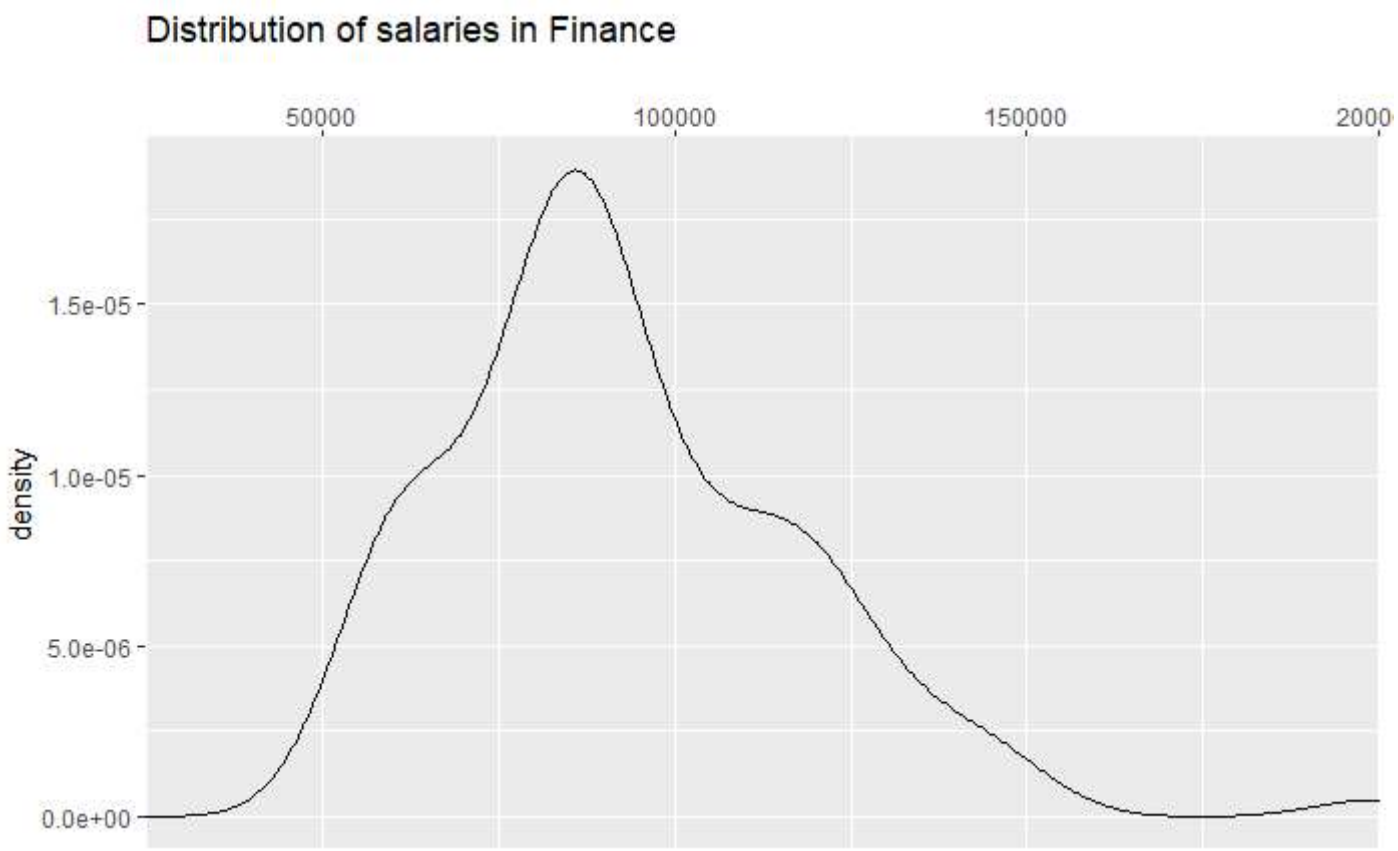
1 row

En data hay 139 posiciones el 90 % de las posiciones tienen menos de 1600 puestos disponibles y el 80 % menos de 860 posiciones, este datos es fundamental para definir la oferta disponible de puestos. Solo 2 vacantes tienen una demanda para mas de 10 mil candidatos df2d %>% arrange(desc(positions_available)) %>% head(20)

```
grouped_df [139 × 5] (S3: grouped_df/tbl_df/tbl/data.frame)
 $ id                 :integer64 [1:139] 16 26 40 50 114 128 156 178 ...
 $ job_title.x        : chr [1:139] "analyst" "data scientist" "data analyst" "research scientist" ...
 $ positions_available: num [1:139] 23681 14333 9223 7912 4295 ...
 $ avg_salary         : num [1:139] 79724 114147 78578 101973 56669 ...
 $ cat                : chr [1:139] "data" "data" "data" "data" ...
 - attr(*, "groups")= tibble [1 × 2] (S3: tbl_df/tbl/data.frame)
  ..$ cat  : chr "data"
  ..$ .rows: list<int> [1:1]
  .. ..$ : int [1:139] 1 2 3 4 5 6 7 8 9 10 ...
  .. ..@ ptype: int(0)
  ..- attr(*, ".drop")= logi TRUE
```

cat <chr>	mean <dbl>	median <dbl>	iqr <dbl>	quantile <dbl>	max <dbl>	sd <dbl>
finance	92853.96	86926.5	31028.25	78332.25	198961	25225.41

1 row



cat <chr>	quantile(avg_salary, 0.9) <dbl>
finance	124696.8

1 row

En finance los salarios se encuentran concentrados cerca de la media de $92000 anuales La mayoria de las posiciones por encima de la media son de posiciones senior, y como posiciones destacadasd las de CFO.

id <S3: integer64>	job_title.x <chr>	positions_available <dbl>	avg_salary <dbl>	cat <chr>
624	chief financial officer	644	198961	finance
1445	tax senior manager	216	149296	finance

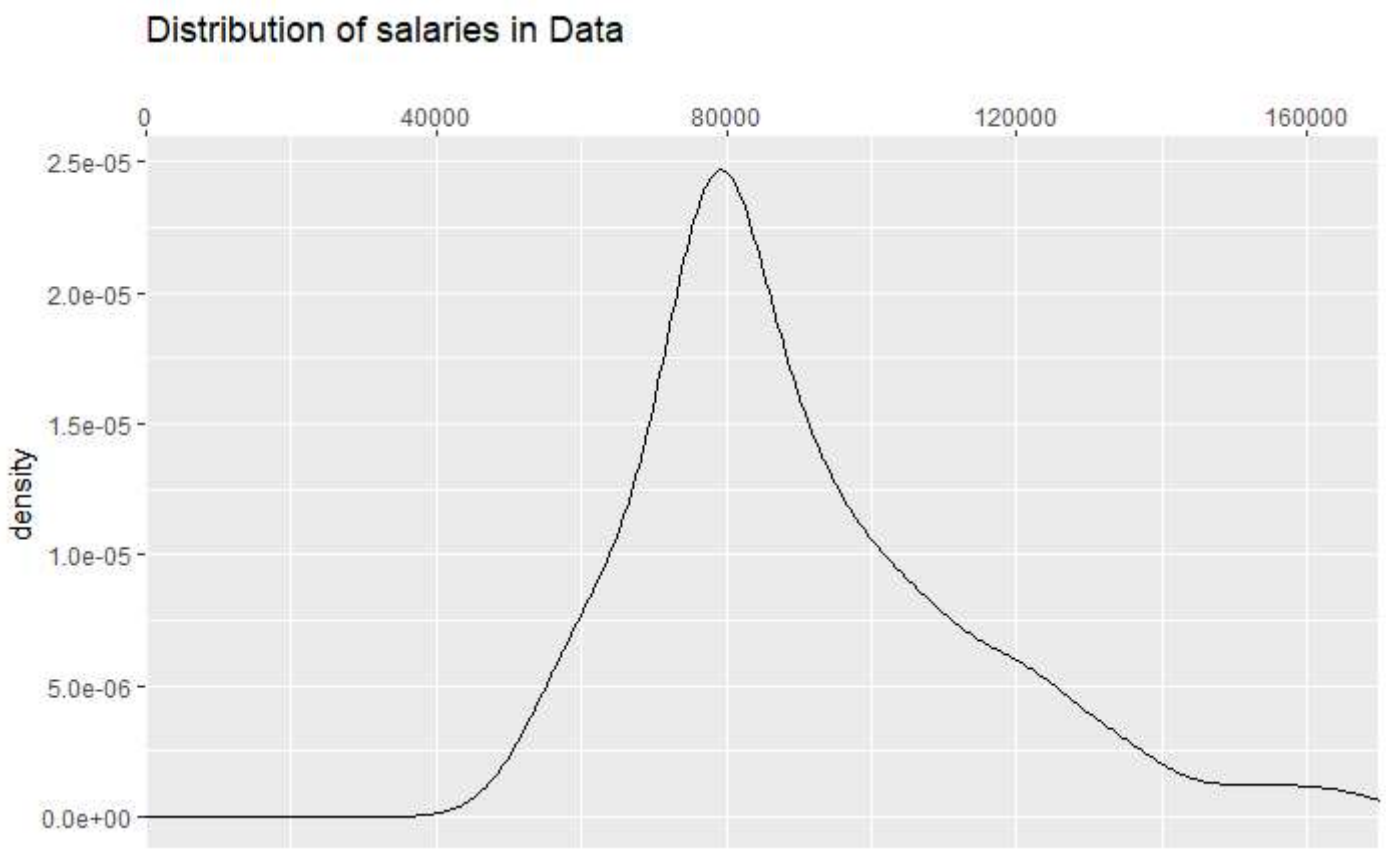
id	job_title.x	positions_available	avg_salary	cat
<S3: integer64>	<chr>	<dbl>	<dbl>	<chr>
1688	business strategy manager	180	144449	finance
1587	business processes principal consultant	195	143842	finance
1293	principal - business consulting	248	140479	finance
1819	business operations consultant	164	133382	finance
1675	senior finance manager	182	132759	finance
1103	finance associate	302	130625	finance
891	business program manager	408	125335	finance
1138	business processes senior consultant	289	124918	finance

1-10 of 98 rows

Previous **1** 2 3 4 5 6 ... 10 Next

cat	mean	median	iqr	quantile	max	sd
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
data	90203.12	82758	25589.5	76841.5	166346	22730.93

1 row



cat <chr>	quantile(avg_salary, 0.9) <dbl>
finance	124696.8

1 row

En data los salarios se encuentran concentrados cerca de la media de $82000 anuales La mayoria de las posiciones por encima de la media son de posiciones senior, y con las 10 posiciones top tienen salarios superiores a $127 mil anuales.

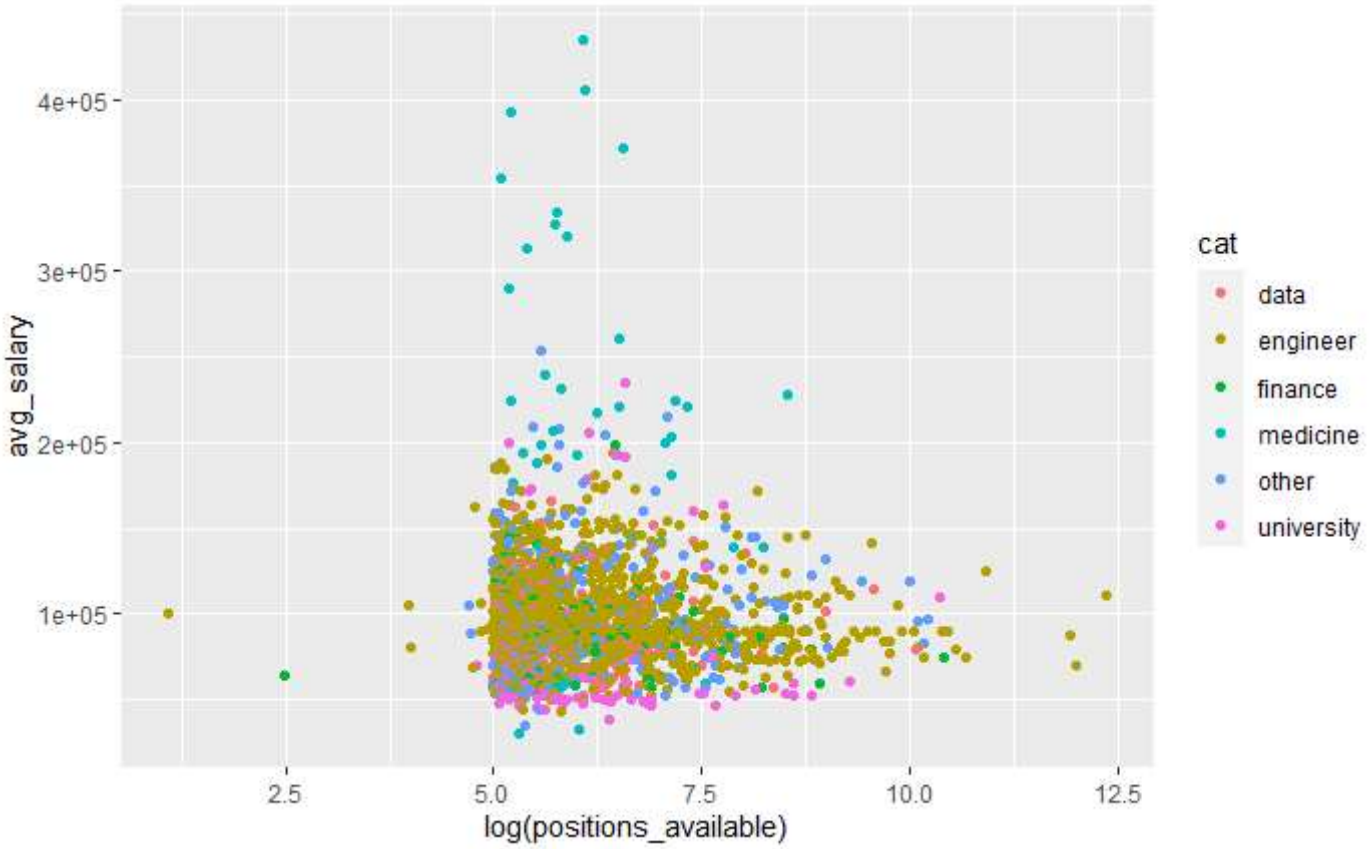
id <S3: integer64>	job_title.x <chr>	positions_available <dbl>	avg_salary <dbl>
1120	data science manager	295	166346
1593	applied scientist iii	193	162453

id <S3: integer64>	job_title.x <chr>	positions_available <dbl>	avg_salary <dbl>
1245	principal data scientist	259	153075
411	applied scientist ii	1014	152309
276	data and applied scientist	1615	142414
665	applied scientist i	595	136235
156	senior data scientist	3034	135579
1126	applied scientist	293	134197
879	lead data scientist	418	132654
1836	machine learning scientist	162	128152

1-10 of 139 rows | 1-4 of 5 columns

Previous **1** 2 3 4 5 6 … 14 Next

Existe correlacion entre las posiciones disponibles y el salario promedio ofrecido?




```
[1] -0.02279816
```

	id	job_title.x	positions_available	avg_salary	cat
	<S3: integer64>	<chr>	<dbl>	<dbl>	<chr>
1	1234	partner	261	254042	other
2	359	chief executive officer	1192	214977	other
3	1346	senior vice president	236	208825	other
4	1039	president	327	208139	other
5	689	principal	567	204629	other
6	1038	internist	327	198961	other

	id	job_title.x	positions_available	avg_salary	cat
	<S3: integer64>	<chr>	<dbl>	<dbl>	<chr>
7	1055	principal program manager	320	185898	other
8	1947	associate partner	151	185368	other
9	847	principal product manager	433	176597	other
10	1661	ceo	184	172241	other

1-10 of 40 rows Previous **1** 2 3 4 Next

En terminos generales, no existe correlacion alguna entre la cantidad de puestos demandados y los salarios ofrecidos, esto en algunos casos tiene una logica, ya que hay puestos como los de cat Others, vinculados a posiciones de management de elevado nivel como CEO que tienen menor demanda en cuanto su oferta, pero el nivel de salarios ofrecidos debe ser elevado. Pero en el caso de posiciones medias, existen posiciones con una alta demanda que poseen una oferta salarial actual promedio, igual que otros puestos con menor demanda lo que nos lleva a pensar que talvez exista cierta cartelizacion en la fijacion de salarios. Otro motivo puede ser que esta desviacion en la curva de oferta y demanda laboral, tienda a corregirse, a medida de que transcurra el tiempo. Es un buen indicio para desarrollar expertise en esas areas.

engineer categorias mas demandadas

	id	job_title.x	positions_available	avg_salary	cat
	<S3: integer64>	<chr>	<dbl>	<dbl>	<chr>
1	1	software engineer	234358	111361	engineer
2	2	programmer analyst	163226	69528	engineer
3	3	software developer	149795	88065	engineer
4	4	senior software engineer	55124	125289	engineer
5	5	systems analyst	43107	74656	engineer
6	6	senior systems analyst jc60	38372	79523	engineer
7	7	computer programmer	34945	90431	engineer
8	9	developer	32082	90431	engineer
9	13	computer systems analyst	26117	75187	engineer
10	15	technology lead - us	23959	80803	engineer

1-10 of 40 rows Previous **1** 2 3 4 Next

Viendo el avg salary de categorias senior

	id	job_title.x	positions_available	avg_salary	cat
	<S3: integer64>	<chr>	<dbl>	<dbl>	<chr>
1	4	senior software engineer	55124	125289	engineer
2	6	senior systems analyst jc60	38372	79523	engineer
3	14	senior consultant	24341	95692	other
4	20	senior software developer	19054	104663	engineer
5	65	senior engineer	6499	111232	engineer
6	73	senior manager jc45	5857	111536	other
7	89	senior systems analyst	5186	88424	engineer
8	99	senior developer	4867	108811	engineer
9	100	senior business analyst	4809	97067	finance
10	103	advisory senior	4782	105285	other

1-10 of 20 rows Previous **1** 2 Next

mean(avg_salary)	median(avg_salary)
<dbl>	<dbl>
107972	105085

1 row

Si comparamos el avg de salarios general contra el que se corresponde con posiciones senior si observamos una clara tendencia a un mayor nivel de salario para las posiciones que tienen esta denominacion.

mean(avg_salary)	median(avg_salary)
<dbl>	<dbl>
98761.81	90859

1 row

ver cuales son las posiciones mas interesantes si consideramos su demanda y su salario.

cat <chr>	mean(avg_salary) <dbl>	median <dbl>	IQR(avg_salary) <dbl>	quantile(avg_salary, 0.25) <dbl>
medicine	185950.60	190413.0	127475.50	101067.25
engineer	97959.43	91004.0	27062.00	83394.50
other	99971.93	90922.0	38835.50	78299.50
finance	92853.96	86926.5	31028.25	78332.25
data	90203.12	82758.0	25589.50	76841.50
university	80399.61	63001.0	44497.50	52132.00

6 rows

cat <chr>	mean <dbl>	median <dbl>	iqr <dbl>	quantile <dbl>	max <dbl>	sd <dbl>
university	1086.8522	344.0	487.50	222.00	31798	3228.4102
medicine	707.5577	334.5	497.50	219.00	5028	917.9621
engineer	1592.9187	332.0	562.00	207.00	234358	9577.9313
finance	1099.2347	322.0	470.25	218.25	33275	3519.0789
other	1106.5325	311.0	575.00	193.75	27533	3065.4343
data	947.3309	295.0	545.00	187.00	23681	2541.8925

6 rows

id <S3: integer64>	job_title.x <chr>	positions_available <dbl>	avg_salary <dbl>	cat <chr>
1	software engineer	234358	111361	engineer

id	job_title.x	positions_available	avg_salary	cat
<S3: integer64>	<chr>	<dbl>	<dbl>	<chr>
4	senior software engineer	55124	125289	engineer
10	assistant professor	31798	109364	university
18	associate	21741	119465	other
20	senior software developer	19054	104663	engineer
26	data scientist	14333	114147	data
27	software development engineer ii	13975	141675	engineer
29	manager	12446	119209	other
32	computer systems analysts	10801	111461	engineer
36	software development engineer i	9595	114197	engineer

1-10 of 95 rows Previous **1** 2 3 4 5 6 … 10 Next

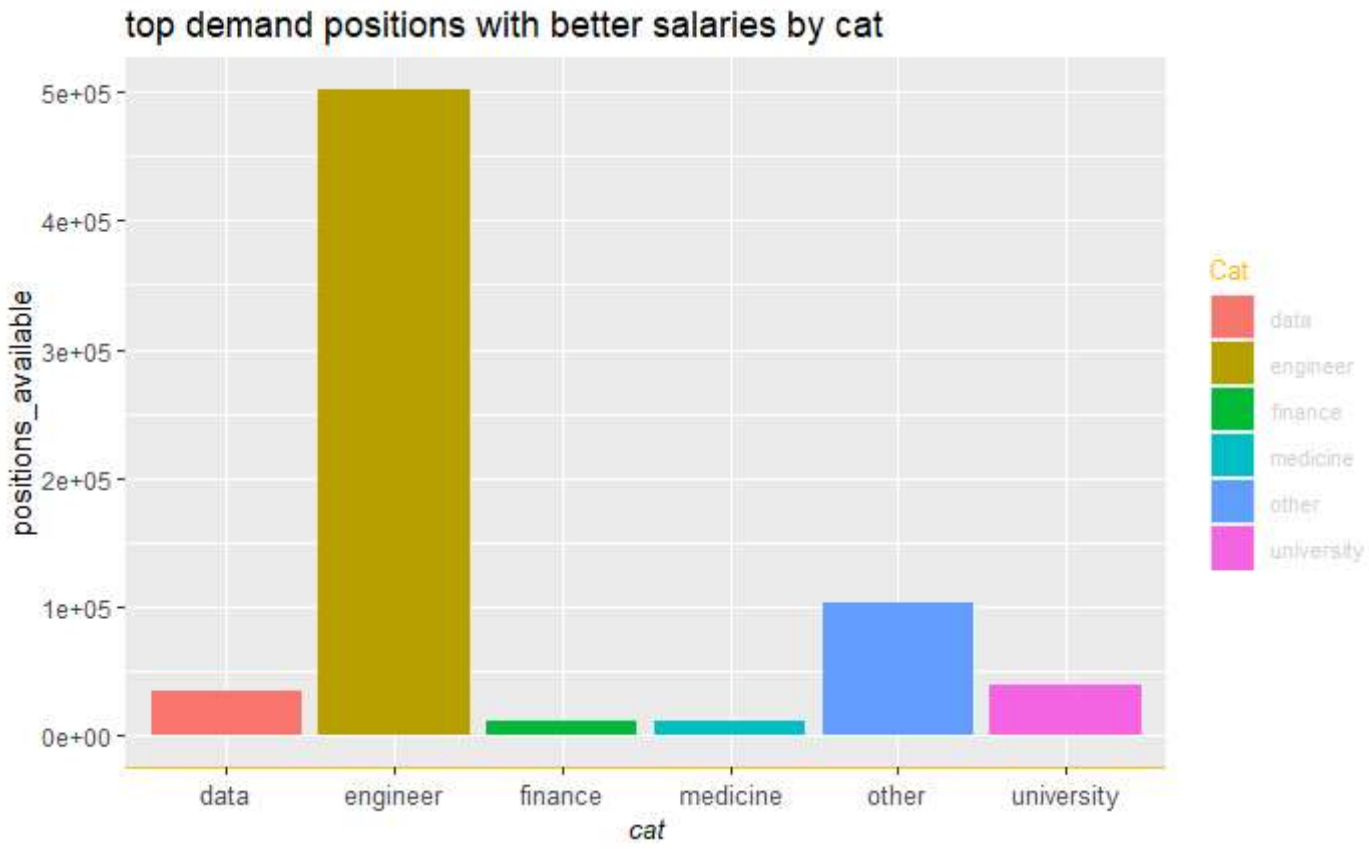
```
grouped_df [95 × 5] (S3: grouped_df/tbl_df/tbl/data.frame)
 $ id                :integer64 [1:95] 1 4 10 18 20 26 27 29 ...
 $ job_title.x       : chr [1:95] "software engineer" "senior software engineer" "assistant professor" "associate" ...
 $ positions_available: num [1:95] 234358 55124 31798 21741 19054 ...
 $ avg_salary        : num [1:95] 111361 125289 109364 119465 104663 ...
 $ cat               : chr [1:95] "engineer" "engineer" "university" "other" ...
 - attr(*, "groups")= tibble [6 × 2] (S3: tbl_df/tbl/data.frame)
  ..$ cat  : chr [1:6] "data" "engineer" "finance" "medicine" ...
  ..$ .rows: list<int> [1:6]
  .. ..$ : int [1:9] 6 15 45 54 77 79 85 93 95
  .. ..$ : int [1:48] 1 2 5 7 9 10 11 12 14 16 ...
  .. ..$ : int [1:5] 29 78 82 83 88
  .. ..$ : int [1:5] 26 80 84 87 94
  .. ..$ : int [1:23] 4 8 13 17 21 31 32 33 35 39 ...
  .. ..$ : int [1:5] 3 56 68 76 86
  .. ..@ ptype: int(0)
  ..- attr(*, ".drop")= logi TRUE
```

cat <chr>	positions <dbl>	mean(avg_salary) <dbl>	n() <int>	percent <dbl>
data	34398	118506.2	9	9.473684
engineer	501672	118701.4	48	50.526316
finance	10474	99438.2	5	5.263158
medicine	10273	214996.4	5	5.263158
other	102821	124389.0	23	24.210526
university	38907	134115.6	5	5.263158

6 rows

Podemos considerar que existen menos de 100 posiciones que superan el promedio en salarios y en demanda al mismo tiempo, de las cuales, el 50.5% son posiciones de ingenieria y alrededor del 10% pertenecen a data. La mas demandada en este top filtro es data scientist que podemos considerarla la mejor oferta en este rango.

id <S3: integer64>	job_title.x <chr>	positions_available <dbl>	avg_salary <dbl>	cat <chr>
411	applied scientist ii	1014	152309	data
276	data and applied scientist	1615	142414	data
156	senior data scientist	3034	135579	data
368	data scientist ii	1169	122282	data
26	data scientist	14333	114147	data
313	tax manager	1394	109758	finance
271	management analysts	1634	107596	data
50	research scientist	7912	101973	data
272	business development manager	1632	101283	finance
100	senior business analyst	4809	97067	finance



Analyzing General Salaries Overview: US Market

manteniendo la tendencia general, las posiciones mas demandandas y con los salarios mas interesantes predominan en el sector de ingenieria e IT.

Summary

Podemos centrarnos en posiciones de IT con alta demanda para garantizar que nuestros clientes puedan acceder a estos puestos antes que la competencia, pero una buena estrategia para garantizar un flujo de trabajo continuo, es hacer foco en las posiciones que tienen en cada categoria al menos la mediana, en demanda, para tener un flujo constante. Podemso suponer que los salarios para ciertas actividades tederan a crecer, a medida que pase el tiempo, ya que las posiciones no se cubren.

El sector de medicina es muy interesante para el flujo de ingresos de la empresa, poque tiene las posiciones con los salarios mas altos, aunque su demanda es considerablemente inferior a las posiciones IT, tener una buena base en el sector medicina puede ofrecer a la empresa una ventaja competitiva.