PRA2 - Tipologia y Ciclo de Vida de los Datos

Geovanny Risco y Robert Novak

3/1/2022

Contents

Descripción del Dataset

```
##      index        Job Title         Salary Estimate    Job Description
##  Min.   :  0.0   Length:672         Length:672         Length:672
##  1st Qu.:167.8   Class :character   Class :character   Class :character
##  Median :335.5   Mode  :character   Mode  :character   Mode  :character
##  Mean   :335.5
##  3rd Qu.:503.2
##  Max.   :671.0
##      Rating       Company Name         Location         Headquarters
##  Min.   :-1.000   Length:672         Length:672         Length:672
##  1st Qu.: 3.300   Class :character   Class :character   Class :character
##  Median : 3.800   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 3.519
##  3rd Qu.: 4.300
##  Max.   : 5.000
##      Size              Founded     Type of ownership    Industry
##  Length:672         Min.   :  -1   Length:672         Length:672
##  Class :character   1st Qu.:1918   Class :character   Class :character
##  Mode  :character   Median :1995   Mode  :character   Mode  :character
##                     Mean   :1636
##                     3rd Qu.:2009
##                     Max.   :2019
##     Sector            Revenue          Competitors
##  Length:672         Length:672         Length:672
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
```

El dataset que hemos escogido está recogido mediante webscraping en distintas plataformas sobre ofertas de empleo relacionadas con los datos en Estados Unidos. Hemos escogido este dataset principalmente por dos razones.

1. Es un dataset bastante desordenado en el que da lugar a hacer procesos de limpieza de distinto tipo ideal para asentar los conceptos tratados en la asignatura
2. Nos parece interesante conocer el mercado laboral de las distintas profesiones a las que podríamos aspirar tras la finalización del máster y la demanda, aunque sea en un país extranjero.

Nuestro principal objetivo con el dataset es contestar a distintas preguntas relacionadas con el salario y distintas variables que se proporcionan en el dataset:

Integración y limpieza de datos

En primer lugar, hemos realizado un análisis del dominio de las variables a partir de las cuáles hemos hecho hecho las siguientes observaciones:

1. Se utiliza el `-1` para indicar valores faltantes. Adicionalmente, existen columnas que tienen un valor faltante que se representa de forma distinta a `-1` por la forma en la que se han extraído los datos. En la limpieza hemos tratado todos esos casos y representado los valores faltantes de forma homogénea mediante `NA`, que es la forma de representar los valores faltantes en R y gracias al cuál podemos hacer operaciones para algunas de las funciones donde se tienen en cuenta los valores faltantes.
2. La columna `Job title` tiene una gran diversidad de trabajos con una mínima variación en la que es interesante tratarlos como un mismo trabajo. Para ello, hemos definido un subconjunto de trabajos a

partir del cuál tratar como iguales las variantes. Ese subconjunto son los que consideramos principales : { data scientist, data engineer, data analyst, machine learning}. Así, por ejemplo, un trabajo de e-commerce data analyst o uno de RFP data analyst será tratado bajo la categoría de data analyst. Aquellos trabajos muy específicos en los cuáles no se engloba bajo ninguna de las categorías anteriores los consideramos muy específicos y, al no ser un número muy elevado hemos decidido eliminarlos del dataset.

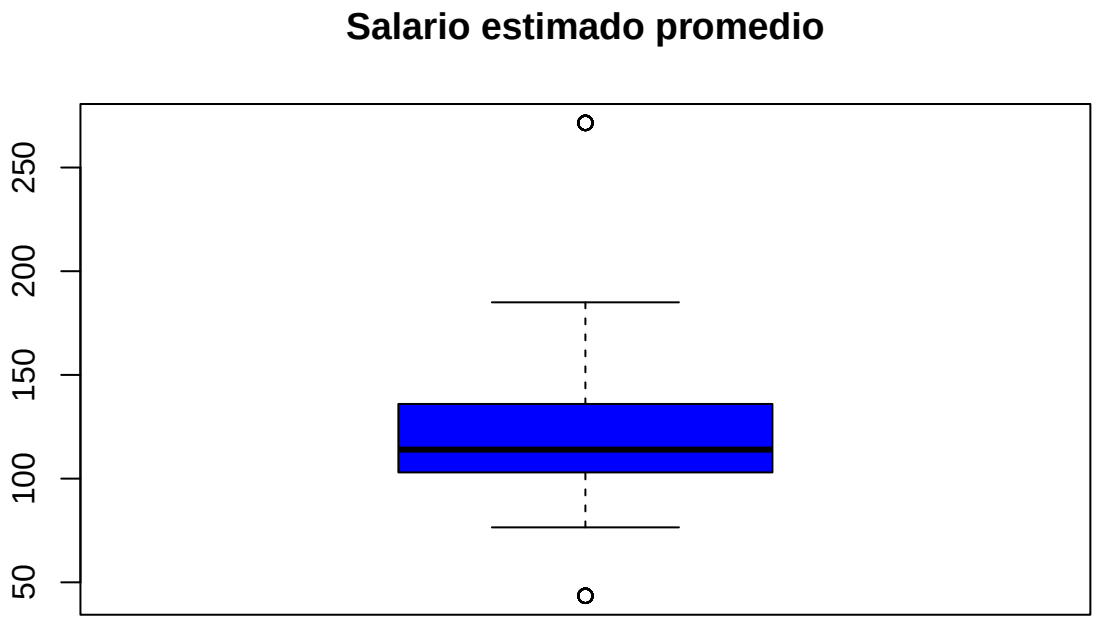
3. La variable `Company name` tiene la información del rating. Hemos eliminado esa redundancia
4. Hemos añadido una nueva variable binaria a partir de `Location` y `Headquarters` para ver aquellas ofertas de trabajo en la que la cede central de la empresa está en el mismo sitio que la oferta
5. Algunas variables como `Salary Estimate`, `Size` y `Revenue` contienen información que pueden ser aprovechadas mejor separándolas en más columnas a partir de las cuáles sacar más información. Así, las hemos separado en más columnas. Una para los rangos mínimos, otro para los rangos máximos y otra para los medios.
6. `Salary Estimate` puede ser considerada una variable cuantitativa ya que, aunque se proporcione un rango variable para todas las ofertas, la realidad es que el salario no es un rango sino un valor concreto dado por un dominio continuo. La decisión que hemos tomado para solucionar esto es considerar el punto medio del rango proporcionado como el salario de la oferta. Esta solución es una aproximación ya que dos ofertas con mismos rangos tendrían el mismo salario y no tendría por qué ser considerados como el mismo. O, incluso, dos salarios con rangos distintos pero con una cierta intersección podrían tener en la realidad el mismo salario pero no tal como lo hemos tratado. Sin embargo, aunque lo ideal sería hacer un estudio externo sobre la distribución del salario dado el rango, la empresa particular, etc. Al no disponer de esa información asumimos esta simplificación.
7. `Size` y `Revenue` deben ser consideradas para análisis posteriores como variables ordinales ya que su dominio corresponde a categorías no solapadas en el que el orden importa.
8. La variable `Job Description` es una variable muy interesante a partir de la cuál se puede obtener información interesante como, por ejemplo, los lenguajes de programación exigidos por la oferta, los años de experiencia necesario, etc. Sin embargo, esta información está presente de forma muy variable de observación en observación lo que hace difícil su extracción para los conocimientos que tenemos actualmente (aún no hemos tenido ninguna asignatura de NLP). Por ello, hemos decidido no tratarla más que para comprobar casos atípicos por si hay alguna información interesante que pudiera explicarlos.

Finalmente, hemos decidido quitar algunas de las variables que venían en el dataset original:

- `index`
- `competitors`
- `founded`
- `Company Name`
- `Type of ownership`

Esto es porque no son útiles porque no aportan información con respecto a la variable de mayor interés que es el salario.

Tratamiento de los valores extremos de la variable de interés `Salary Estimate Med`



En el análisis que hemos realizado, hemos observado que la variable `Salary Estimate Med` tiene una distribución de valores bastante consistente, lo cual podemos comprobar dado el pequeño tamaño de su rango intercuartílico. No obstante, vemos como existen puntos fuera de este rango y de los *whiskers*, los conocidos como *outliers*. En concreto, son dos puntos: 43.5 y 271.5.

Explorando los casos en el que el salario es atípicamente bajo hemos encontrado que se corresponde con posiciones en las que no se piden experiencia (lo podemos comprobar por la variable `Job Description`) y, además, se encuentran en ciudades de Estados Unidos poco punteras (tecnologicamente hablando), como por ejemplo "Lincoln", "Arlington", "Saint Paul", etc. Por tanto, podemos concluir que estas variables no se tratan de errores en el dataset, sino de valores totalmente legítimos que debemos tener en cuenta para las pruebas estadísticas posteriores.

En el caso del valor atípico alto podemos observar lo contrario, estas observaciones están asociadas a:

- Posiciones para managers, seniors o PhDs. Es decir, se requiere formación y experiencia para optar a este sueldo
- Ciudad muy punteras en el ambito tecnológico. Ex. "Seattle", "New York", "Washington", etc.
- Empresas muy top. Ex. "Roche", "Aztrazeneca", "Maxar Technologies", etc.

Dado que estos salarios tan altos son totalmente justificables (por las razones planteadas, entre otras), hemos mantenido los registros para las pruebas estadísticas posteriores.

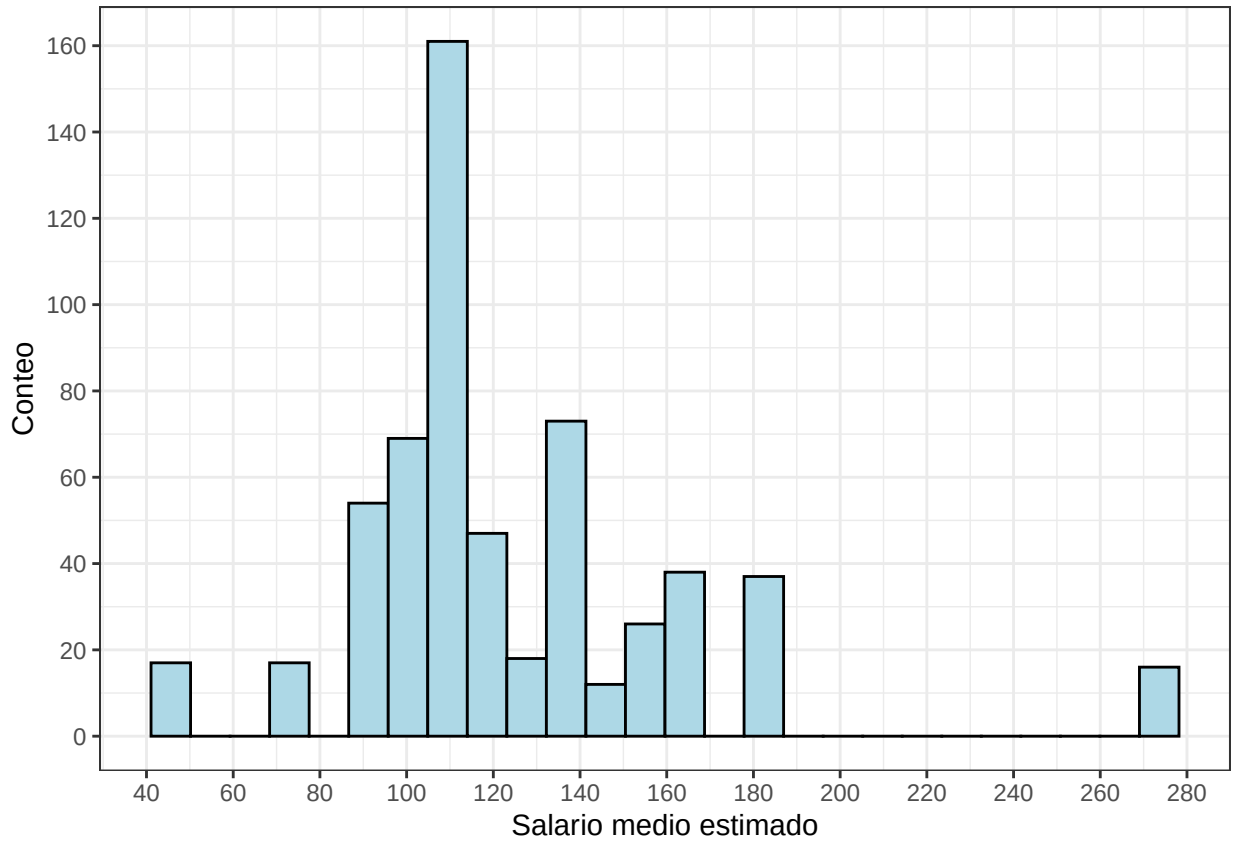
Análisis de datos

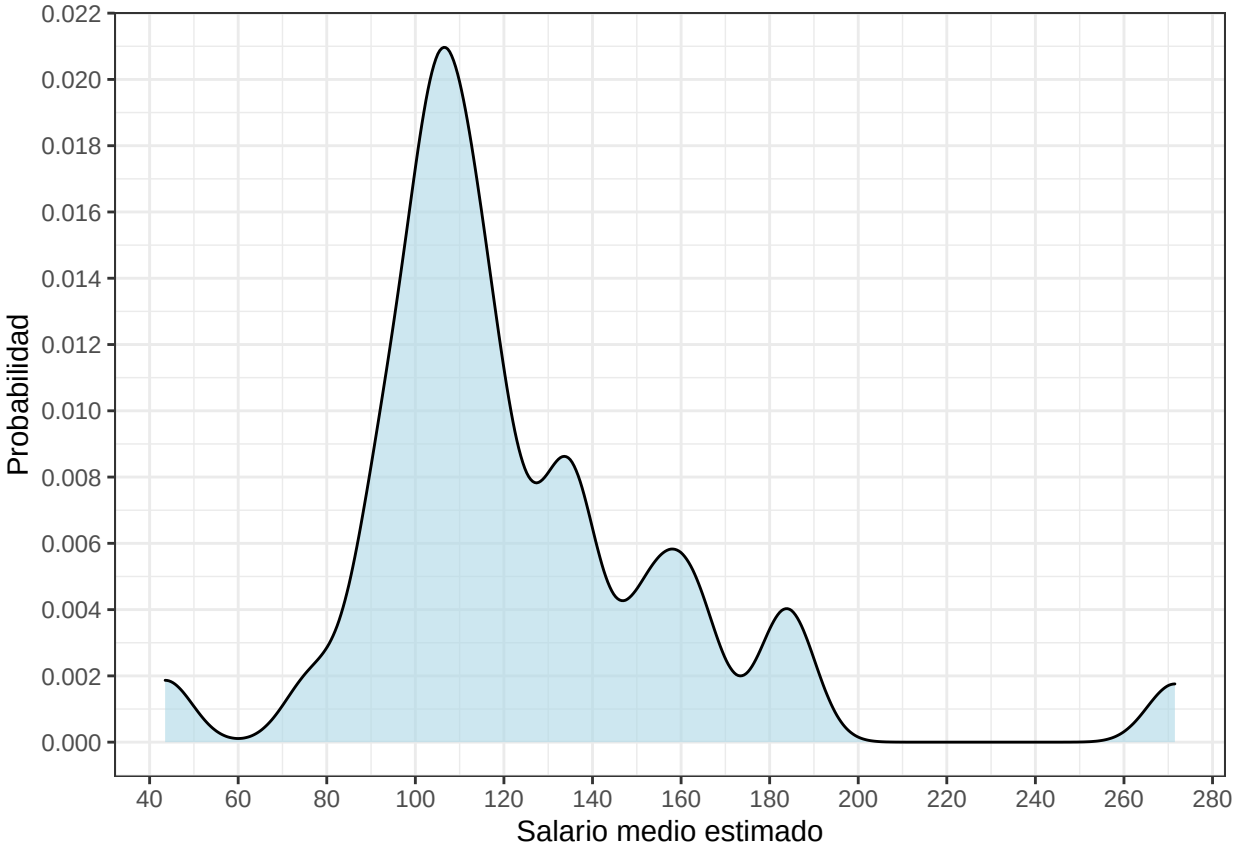
Para el análisis nos interesa la relación existente entre las distintas variables del dataset con la variable `Salary`. Las pruebas estadísticas que se aplicarán dependerán del análisis de la normalidad de `Salary Estimate Med`. Las preguntas de interés son: ¿Existen diferencias salariales significativas entre los distintos trabajos que hemos definido? ¿Existe alguna diferencia respecto a los salarios según el tamaño de la empresa? ¿Y según el revenue? ¿y entre las distintas industrias?

Comprobación de normalidad de la variable `Salary Estimate Med`

La variable `Salary Estimate Med` representa la variable cuantitativa de interés según lo explicado en el apartado de limpieza de datos.

A continuación proporcionamos dos gráficas para comprobar visualmente, previo al test de normalidad, si la variable tiene apariencia de ser normal.





En vista de los gráficos se puede apreciar que no parece que la distribución del salario sea normal. Comprobándolo de manera formal con el test de shapiro, el pvalor es $1.8465838 \times 10^{-22}$. El p valor es muy bajo por lo que se puede rechazar la hipótesis nula de que la variable sigue una distribución normal. Hay que tener en cuenta este hecho para los análisis que presupongan normalidad

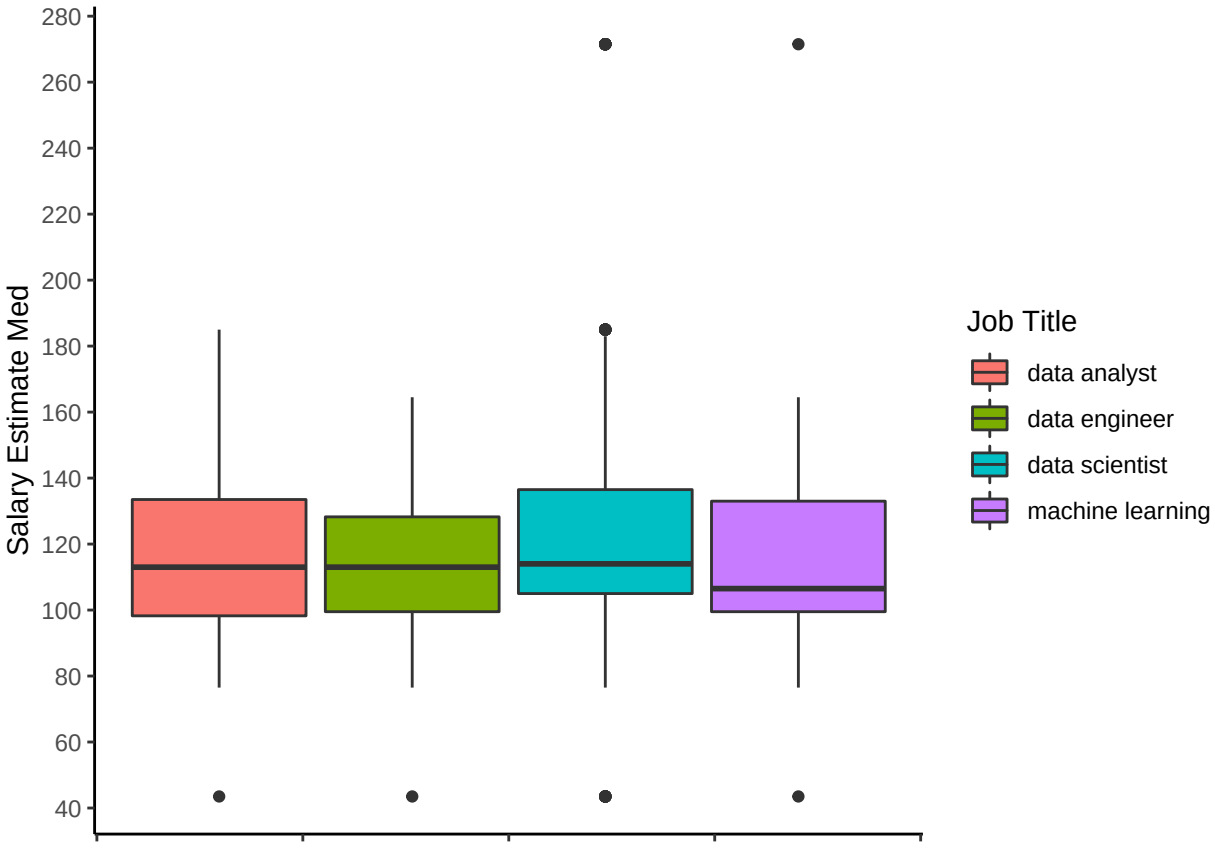
Pruebas estadísticas

Análisis entre las variables categóricas `Job Title`, `Size Ordered`, `Revenue Ordered`, `Industry` y la variable cuantitativa `Salary Estimate Med`

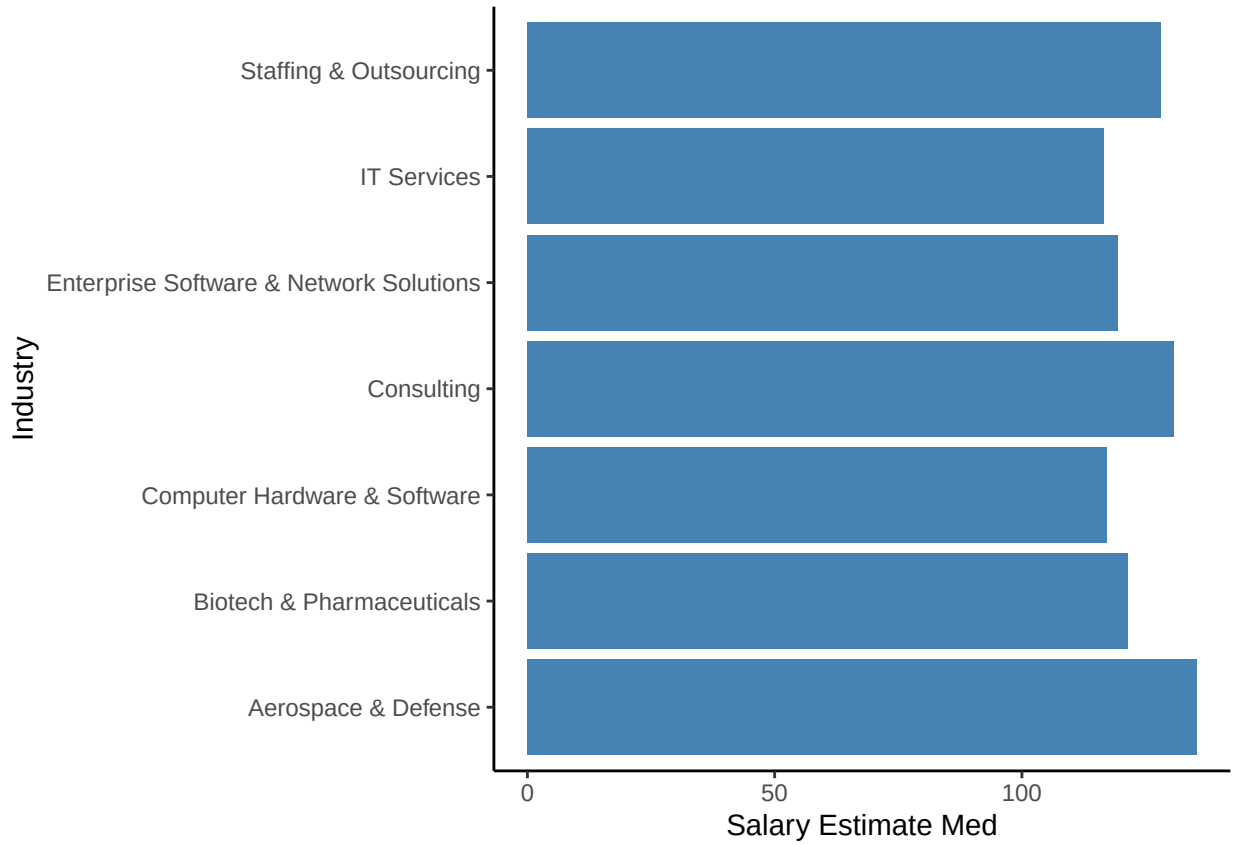
A continuación, vamos a comprobar las diferencias entre los distintos grupos existentes en el dataset.

La cuestión principal es **¿Existen diferencias salariales significativas entre los distintos trabajos que hemos definido?**

Si representamos esto gráficamente, se puede apreciar que los salarios se distribuyen de forma bastante uniforme:



De la misma manera, sería interesante observar las diferencias entre los salarios para cada una de las industrias disponibles en el dataset. Dado que existen más de 50 valores diferentes para la variable de `Industry`, procederemos a representar al menos las 7 primeras industrias con más concurrencia en el dataset.



A simple vista no observamos diferencias considerables entre los salarios medios para cada una de las industrias mostradas en el gráfico. Vemos como casi todas ofrecen salarios cercanos a los 125.000$. La industria aeroespacial es la que mejor pagada, superando los 150.000$ anuales.

Vamos a comprobar lo visto anteriormente de manera formal. Para ello, hemos comrpobado, en primer lugar, la homocedasticidad mediante el test de levene para todas las variables de interés. Los resultados han sido 0.1460238, 0.9684899, 0.9442937 y 0.9409076 para `Job Title`, `Size Ordered`, `Revenue Ordered`, `Industry` respectivamente. Es decir, en todos los casos no podemos rechazar la hipótesis nula de que las varianzas poblacionales son iguales. Por tanto, podemos aplicar el test de Kruskal en el cual no hace falta que las distribuciones provengan de la normal pero sí asume la homocedasticidad.

Job Title	Size	Revenue	Industry
0.1460238	0.9684899	0.9442937	0.9409076

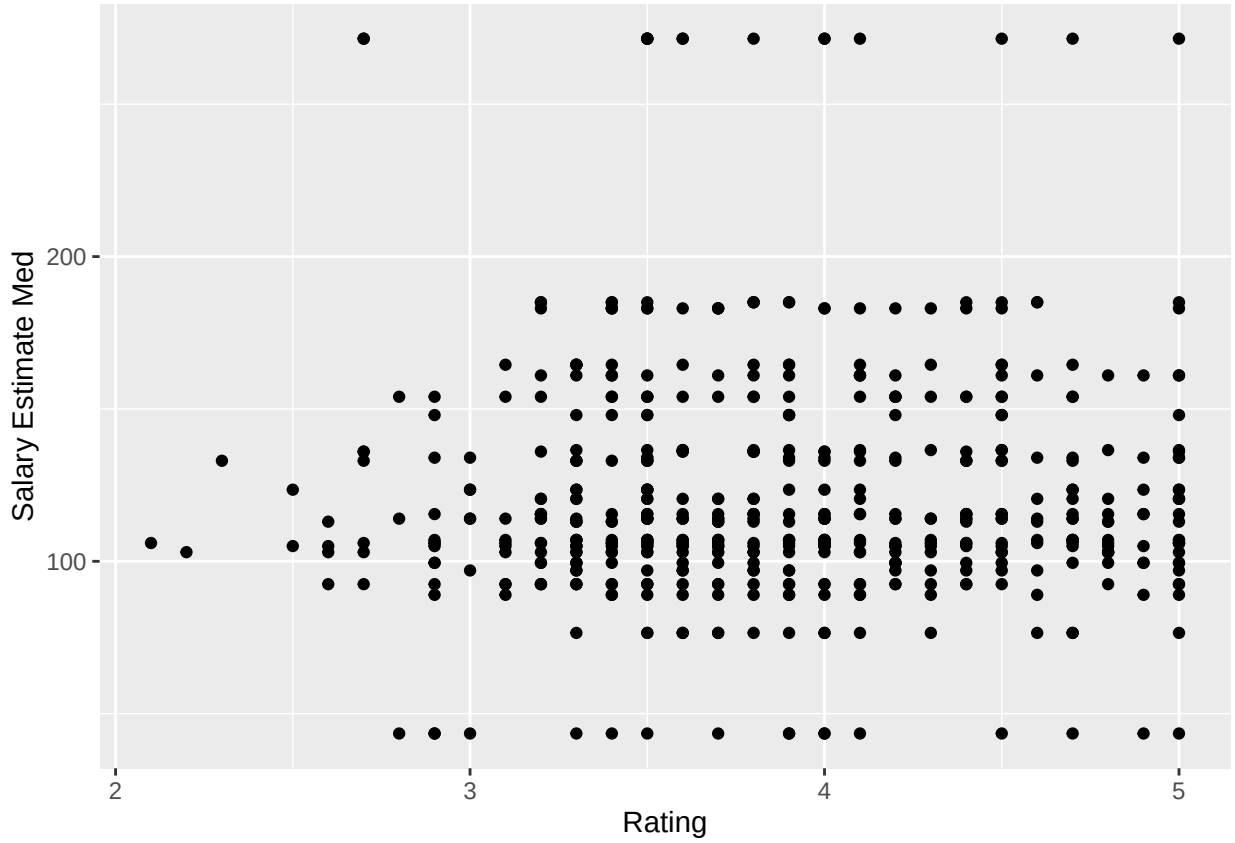
El test de Kruskal es un test no paramétrico cuya hipótesis inicial es que todas las muestras provienen de la misma distribución. Mediante él, hemos comprobado formalmente que esto efectivamente es así, pues los p valores son los mostrados en la tabla resumen de a continuación:

Job Title	Size	Revenue	Industry
0.1105966	0.4131254	0.0912716	0.1325356

Análisis de correlación entre `Rating` y `Salary Estimate Med`

Otra cuestión de interés es si a partir del rating se puede predecir el salario. Para ver si la calidad de la oferta está relacionada con el salario o quizás intervengan más factores. La relación entre las dos variables se puede

ver en el siguiente gráfico:



La correlación de Spearman que nos ha salido entre estas dos variables es de 0.0253103. Es decir, no parece que estén correlacionadas.

Por último, hemos hecho una regresión lineal múltiple con las variables que hemos estudiado, aunque por los análisis previos ya sospechábamos que este modelo no podría explicar la varianza de los datos. Efectivamente, esto se puede ver en el siguiente summary

```
##
## Call:
## lm(formula = `Salary Estimate Med` ~ `Job Title` + as.factor(`Size Ordered`) +
##     as.factor(`Revenue Ordered`) + Industry, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -83.398 -18.866  -5.604  14.512 151.247
##
## Coefficients:
##                                                Estimate Std. Error t value
## (Intercept)                                    126.7386    23.5645   5.378
## `Job Title`data engineer                        -6.8417    11.9118  -0.574
## `Job Title`data scientist                       -2.0720     8.8366  -0.234
## `Job Title`machine learning                     -8.5357    14.4301  -0.592
## as.factor(`Size Ordered`)2                       6.3550    11.9233   0.533
## as.factor(`Size Ordered`)3                       6.6597    12.8845   0.517
## as.factor(`Size Ordered`)4                      -0.7779    15.2268  -0.051
## as.factor(`Size Ordered`)5                      -2.2661    14.5294  -0.156
```

```
## as.factor(`Size Ordered`)6                          -2.0848    17.6081  -0.118
## as.factor(`Size Ordered`)7                          -1.7816    19.7512  -0.090
## as.factor(`Revenue Ordered`)2                       11.7745    19.1796   0.614
## as.factor(`Revenue Ordered`)3                        5.3136    22.2896   0.238
## as.factor(`Revenue Ordered`)4                       13.2167    20.0953   0.658
## as.factor(`Revenue Ordered`)5                        9.6313    21.1596   0.455
## as.factor(`Revenue Ordered`)6                      -14.0194    22.0132  -0.637
## as.factor(`Revenue Ordered`)7                       13.8285    21.3380   0.648
## as.factor(`Revenue Ordered`)8                       26.5512    27.0040   0.983
## as.factor(`Revenue Ordered`)9                        5.3036    24.8302   0.214
## as.factor(`Revenue Ordered`)10                      17.8644    24.4152   0.732
## as.factor(`Revenue Ordered`)11                      28.4628    30.5566   0.931
## as.factor(`Revenue Ordered`)12                       9.6237    27.1114   0.355
## IndustryAdvertising & Marketing                    -22.7205    30.0122  -0.757
## IndustryAerospace & Defense                         -3.7520    29.5185  -0.127
## IndustryArchitectural & Engineering Services        19.9558    36.1365   0.552
## IndustryBanks & Credit Unions                      -10.5008    38.1256  -0.275
## IndustryBiotech & Pharmaceuticals                  -12.2555    29.6431  -0.413
## IndustryCable, Internet & Telephone Providers      -33.0088    47.9774  -0.688
## IndustryChemical Manufacturing                     -33.7017    42.5672  -0.792
## IndustryComputer Hardware & Software               -18.2568    28.7823  -0.634
## IndustryConstruction                               -61.4791    39.5374  -1.555
## IndustryConsulting                                  -3.5131    29.3106  -0.120
## IndustryConsumer Electronics & Appliances Stores     1.7958    41.0259   0.044
## IndustryConsumer Products Manufacturing            -43.7269    37.1219  -1.178
## IndustryDepartment, Clothing, & Shoe Stores        -25.5088    47.9774  -0.532
## IndustryElectrical & Electronic Manufacturing        9.9796    36.2317   0.275
## IndustryEnergy                                     -31.9517    39.6657  -0.806
## IndustryEnterprise Software & Network Solutions    -15.8963    28.4463  -0.559
## IndustryFederal Agencies                            7.4756    31.4430   0.238
## IndustryFinancial Transaction Processing           -28.5506    39.2706  -0.727
## IndustryFood & Beverage Manufacturing               -9.5088    35.0446  -0.271
## IndustryFood & Beverage Stores                      45.2827    47.9582   0.944
## IndustryHealth Care Services & Hospitals           -21.6685    30.1889  -0.718
## IndustryHealth, Beauty, & Fitness                   67.4148    40.9180   1.648
## IndustryIndustrial Manufacturing                    -7.3496    36.3429  -0.202
## IndustryInsurance Agencies & Brokerages            -41.9779    34.9253  -1.202
## IndustryInsurance Carriers                         -24.5366    30.4745  -0.805
## IndustryInternet                                    -5.5787    31.0988  -0.179
## IndustryInvestment Banking & Asset Management      -25.9371    32.9663  -0.787
## IndustryIT Services                                -23.4703    28.6575  -0.819
## IndustryLending                                     -4.5884    33.5013  -0.137
## IndustryLogistics & Supply Chain                   -34.7495    48.9181  -0.710
## IndustryNews Outlet                                -10.3070    48.2255  -0.214
## IndustryOil & Gas Services                         -51.3789    40.9586  -1.254
## IndustryOther Retail Stores                         45.2827    47.9582   0.944
## IndustryRail                                       -26.2650    48.6693  -0.540
## IndustryReal Estate                                -30.2411    39.6320  -0.763
## IndustryResearch & Development                       4.9800    32.2851   0.154
## IndustrySocial Assistance                          -97.1529    47.9258  -2.027
## IndustryStaffing & Outsourcing                     -25.2791    29.1883  -0.866
## IndustryState & Regional Agencies                   50.4912    47.9774   1.052
## IndustryTelecommunications Manufacturing           -31.1887    49.0474  -0.636
## IndustryTelecommunications Services                -18.2327    34.6367  -0.526
```

```
## IndustryTimber Operations                          -25.4387    41.1078  -0.619
## IndustryTransportation Equipment Manufacturing      31.6521    51.2125   0.618
## IndustryTransportation Management                  -25.5088    47.9774  -0.532
## IndustryUtilities                                  -28.5150    40.6558  -0.701
## IndustryVenture Capital & Private Equity             2.4628    40.2172   0.061
## IndustryVideo Games                                 55.2958    48.9788   1.129
## IndustryWholesale                                   76.0458    41.0259   1.854
##                                                   Pr(>|t|)
## (Intercept)                                       1.56e-07 ***
## `Job Title`data engineer                             0.5662
## `Job Title`data scientist                            0.8148
## `Job Title`machine learning                          0.5546
## as.factor(`Size Ordered`)2                           0.5945
## as.factor(`Size Ordered`)3                           0.6056
## as.factor(`Size Ordered`)4                           0.9593
## as.factor(`Size Ordered`)5                           0.8762
## as.factor(`Size Ordered`)6                           0.9058
## as.factor(`Size Ordered`)7                           0.9282
## as.factor(`Revenue Ordered`)2                        0.5398
## as.factor(`Revenue Ordered`)3                        0.8117
## as.factor(`Revenue Ordered`)4                        0.5113
## as.factor(`Revenue Ordered`)5                        0.6493
## as.factor(`Revenue Ordered`)6                        0.5247
## as.factor(`Revenue Ordered`)7                        0.5175
## as.factor(`Revenue Ordered`)8                        0.3263
## as.factor(`Revenue Ordered`)9                        0.8310
## as.factor(`Revenue Ordered`)10                       0.4650
## as.factor(`Revenue Ordered`)11                       0.3524
## as.factor(`Revenue Ordered`)12                       0.7229
## IndustryAdvertising & Marketing                      0.4496
## IndustryAerospace & Defense                          0.8989
## IndustryArchitectural & Engineering Services         0.5812
## IndustryBanks & Credit Unions                        0.7832
## IndustryBiotech & Pharmaceuticals                    0.6796
## IndustryCable, Internet & Telephone Providers        0.4920
## IndustryChemical Manufacturing                       0.4292
## IndustryComputer Hardware & Software                 0.5264
## IndustryConstruction                                 0.1211
## IndustryConsulting                                   0.9047
## IndustryConsumer Electronics & Appliances Stores     0.9651
## IndustryConsumer Products Manufacturing              0.2398
## IndustryDepartment, Clothing, & Shoe Stores          0.5954
## IndustryElectrical & Electronic Manufacturing        0.7832
## IndustryEnergy                                       0.4212
## IndustryEnterprise Software & Network Solutions      0.5767
## IndustryFederal Agencies                             0.8122
## IndustryFinancial Transaction Processing             0.4678
## IndustryFood & Beverage Manufacturing                0.7863
## IndustryFood & Beverage Stores                       0.3459
## IndustryHealth Care Services & Hospitals             0.4735
## IndustryHealth, Beauty, & Fitness                    0.1005
## IndustryIndustrial Manufacturing                     0.8399
## IndustryInsurance Agencies & Brokerages              0.2304
## IndustryInsurance Carriers                           0.4214
```

```
## IndustryInternet                                    0.8578
## IndustryInvestment Banking & Asset Management      0.4321
## IndustryIT Services                                 0.4135
## IndustryLending                                     0.8912
## IndustryLogistics & Supply Chain                    0.4781
## IndustryNews Outlet                                 0.8309
## IndustryOil & Gas Services                          0.2107
## IndustryOther Retail Stores                         0.3459
## IndustryRail                                        0.5898
## IndustryReal Estate                                 0.4461
## IndustryResearch & Development                      0.8775
## IndustrySocial Assistance                           0.0436 *
## IndustryStaffing & Outsourcing                      0.3872
## IndustryState & Regional Agencies                   0.2935
## IndustryTelecommunications Manufacturing            0.5254
## IndustryTelecommunications Services                 0.5990
## IndustryTimber Operations                           0.5365
## IndustryTransportation Equipment Manufacturing      0.5370
## IndustryTransportation Management                   0.5954
## IndustryUtilities                                   0.4836
## IndustryVenture Capital & Private Equity            0.9512
## IndustryVideo Games                                 0.2598
## IndustryWholesale                                   0.0648 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.84 on 288 degrees of freedom
##   (228 observations deleted due to missingness)
## Multiple R-squared:  0.224,  Adjusted R-squared:  0.04084
## F-statistic: 1.223 on 68 and 288 DF,  p-value: 0.1326
```

Conclusión

En vista del análisis previo concluimos que no existen diferencias signficativas en los salarios para las variables que hemos planificado en el análisis. No hemos encontrado que ninguna de estas variables nos pueda ayudar para predecir el salario. Podría ser interesante como trabajo futuro hacer un análisis parecido pero con la información extraída de la columna `Job Description` donde pueden existir variables más interesantes como, por ejemplo, la experiencia requerida o los lenguajes de programación, entre otras características.

Contribuciones	Firma
Investigación previa	Geovanny Risco, Robert Novak
Redacción de las respuestas	Geovanny Risco, Robert Novak
Desarrollo del código	Geovanny Risco, Robert Novak