Informe

Geovanny Risco y Robert Novak

28/12/2021

Contents

Importación de librerías

```
library(tidyverse)
#library(kableExtra)
library(pander) #Para mostrar los vectores más estéticamente
library(DT)
library(readr)
library(missForest)
```

Importación del dataset

```
raw_data <- read_csv("Data/Raw/uncleaned_data.csv")
```

```
# raw_data %>% DT::datatable(
#   extensions = 'FixedColumns',
#   options = list(
#     dom = 't',
#     scrollX = TRUE,
#     scrollCollapse = TRUE)
#   )
```

Exploración del dominio de las variables

Nombre de todas las variables

```
cols_raw_data <- names(raw_data)
cols_raw_data
```

```
##  [1] "index"             "Job Title"         "Salary Estimate"
##  [4] "Job Description"   "Rating"            "Company Name"
##  [7] "Location"          "Headquarters"      "Size"
## [10] "Founded"           "Type of ownership" "Industry"
## [13] "Sector"            "Revenue"           "Competitors"
```

Selección de las variables de interés para el estudio del dominio

```
cols_no_interest <- c("index","Job Description") #Eliminamos Job Description de momento por tener demas
cols_raw_data_filtered <- cols_raw_data[-which(cols_raw_data %in% cols_no_interest)]
```

Visualización en forma de lista del dominio de cada variable

```
cols_raw_data_filtered %>%
  map(function(x) unique(raw_data[[x]])) %>%
  setNames(cols_raw_data_filtered) -> list_domain
list_domain
```

```
## $`Job Title`
##   [1] "Sr Data Scientist"
##   [2] "Data Scientist"
##   [3] "Data Scientist / Machine Learning Expert"
##   [4] "Staff Data Scientist - Analytics"
```

```
##   [5] "Data Scientist - Statistics, Early Career"
##   [6] "Data Modeler"
##   [7] "Experienced Data Scientist"
##   [8] "Data Scientist - Contract"
##   [9] "Data Analyst II"
##  [10] "Medical Lab Scientist"
##  [11] "Data Scientist/Machine Learning"
##  [12] "Human Factors Scientist"
##  [13] "Business Intelligence Analyst I- Data Insights"
##  [14] "Data Scientist - Risk"
##  [15] "Data Scientist-Human Resources"
##  [16] "Senior Research Statistician- Data Scientist"
##  [17] "Data Engineer"
##  [18] "Associate Data Scientist"
##  [19] "Business Intelligence Analyst"
##  [20] "Senior Analyst/Data Scientist"
##  [21] "Data Analyst"
##  [22] "Machine Learning Engineer"
##  [23] "Data Analyst I"
##  [24] "Scientist - Molecular Biology"
##  [25] "Computational Scientist, Machine Learning"
##  [26] "Senior Data Scientist"
##  [27] "Jr. Data Engineer"
##  [28] "E-Commerce Data Analyst"
##  [29] "Data Analytics Engineer"
##  [30] "Product Data Scientist - Ads Data Science"
##  [31] "Data Scientist - Intermediate"
##  [32] "Global Data Analyst"
##  [33] "Data & Machine Learning Scientist"
##  [34] "Data Scientist - Machine Learning"
##  [35] "Data Engineer (Remote)"
##  [36] "Data Scientist, Applied Machine Learning - Bay Area"
##  [37] "Principal Data Scientist"
##  [38] "Business Data Analyst"
##  [39] "Purification Scientist"
##  [40] "Data Engineer, Enterprise Analytics"
##  [41] "Data Scientist 3 (718)"
##  [42] "Real World Science, Data Scientist"
##  [43] "Data Scientist - Image and Video Analytics"
##  [44] "Data Science Manager, Payment Acceptance - USA"
##  [45] "Data Scientist / Applied Mathematician"
##  [46] "Patient Safety- Associate Data Scientist"
##  [47] "(Sr.) Data Scientist -"
##  [48] "Data Scientist, Kinship - NYC/Portland"
##  [49] "Applied Technology Researcher / Data Scientist"
##  [50] "Health Data Scientist - Biomedical/Biostats"
##  [51] "Staff Data Scientist"
##  [52] "Sr Data Engineer (Sr BI Developer)"
##  [53] "Lead Data Scientist"
##  [54] "RFP Data Analyst"
##  [55] "Data Scientist (TS/SCI)"
##  [56] "Software Engineer - Data Science"
##  [57] "Data Analyst/Engineer"
##  [58] "NGS Scientist"
```

```
##  [59] "Senior Data Engineer"
##  [60] "Sr. ML/Data Scientist - AI/NLP/Chatbot"
##  [61] "Data Integration and Modeling Engineer"
##  [62] "Tableau Data Engineer 20-0117"
##  [63] "AI Data Scientist"
##  [64] "Research Scientist Patient Preferences (Remote)"
##  [65] "Scientist - Biomarker and Flow Cytometry"
##  [66] "Analytics Manager"
##  [67] "Staff Scientist- Upstream PD"
##  [68] "Sr Scientist - Extractables & Leachables"
##  [69] "ELISA RESEARCH SCIENTIST (CV-15)"
##  [70] "Say Business Data Analyst"
##  [71] "Geospatial Data Scientist"
##  [72] "Computational Scientist"
##  [73] "Senior Data Analyst"
##  [74] "Sr Data Analyst"
##  [75] "Machine Learning Scientist - Bay Area, CA"
##  [76] "Senior Data Scientist - Algorithms"
##  [77] "Senior Data & Machine Learning Scientist"
##  [78] "Research Scientist - Patient-Centered Research (Remote)"
##  [79] "Jr. Business Data Analyst (position added 6/12/2020)"
##  [80] "Sr. Data Scientist II"
##  [81] "Production Engineer - Statistics/Data Analysis"
##  [82] "Statistical Scientist"
##  [83] "Computational Behavioral Scientist"
##  [84] "Principal Data Scientist - Machine Learning"
##  [85] "Principal Machine Learning Scientist"
##  [86] "Senior Data Scientist - R&D Oncology"
##  [87] "Health Plan Data Analyst, Sr"
##  [88] "Principal Scientist/Associate Director, Quality Control and Analytical Technologies"
##  [89] "Analytics - Business Assurance Data Analyst"
##  [90] "Senior Data Scientist - Image Analytics, Novartis AI Innovation Lab"
##  [91] "Data Science Instructor"
##  [92] "Senior Business Intelligence Analyst"
##  [93] "In-Line Inspection Data Analyst"
##  [94] "Data Scientist - TS/SCI FSP or CI Required"
##  [95] "Data Scientist - TS/SCI Required"
##  [96] "Data Science Software Engineer"
##  [97] "ENGINEER - COMPUTER SCIENTIST - RESEARCH COMPUTER SCIENTIST - SIGNAL PROCESSING - SAN ANTONIO
##  [98] "AI Ops Data Scientist"
##  [99] "Intelligence Data Analyst, Senior"
## [100] "Analytics Manager - Data Mart"
## [101] "Data Modeler (Analytical Systems)"
## [102] "Senior Machine Learning Scientist - Bay Area, CA"
## [103] "Report Writer-Data Analyst"
## [104] "Staff Data Scientist - Pricing"
## [105] "Equity Data Insights Analyst - Quantitative Analyst"
## [106] "Operations Data Analyst"
## [107] "Software Data Engineer"
## [108] "Real World Evidence (RWE) Scientist"
## [109] "Computer Scientist 1"
## [110] "Environmental Data Science"
## [111] "Staff BI and Data Engineer"
## [112] "Data Scientist - Statistics, Mid-Career"
```

```
## [113] "Director of Data Science"
## [114] "Data Engineer, Digital & Comp Pathology"
## [115] "Manager / Lead, Data Science & Analytics"
## [116] "Diversity and Inclusion Data Analyst"
## [117] "Data Scientist Machine Learning"
## [118] "Chief Scientist"
## [119] "Development Scientist, Voltaren"
## [120] "Principal Data & Analytics Platform Engineer"
## [121] "Machine Learning Engineer/Scientist"
## [122] "Data Analyst - Unilever Prestige"
## [123] "VP, Data Science"
## [124] "Data Engineer - Kafka"
## [125] "Decision Scientist"
## [126] "Data Science All Star Program - Data Engineer Track"
## [127] "Scientist - Machine Learning"
## [128] "Sr. Data Scientist"
## [129] "Applied AI Scientist / Engineer"
## [130] "Data Engineer (Analytics, SQL, Python, AWS)"
## [131] "Senior Data Analyst - Finance & Platform Analytics"
## [132] "Market Research Data Scientist"
## [133] "IT Partner Digital Health Technology and Data Science"
## [134] "Software Engineer (Data Scientist, C,C++,Linux,Unix) - SISW - MG"
## [135] "Senior Clinical Data Scientist Programmer"
## [136] "Computer Vision / Deep Learning Scientist"
## [137] "Data Solutions Engineer - Data Modeler"
## [138] "Data Scientist (TS/SCI w/ Poly)"
## [139] "Weapons and Sensors Engineer/Scientist"
## [140] "Applied Computer Scientist"
## [141] "Cloud Data Engineer (Azure)"
## [142] "Lead Certified Clinical Laboratory Scientist - Saturday - Tuesday, 8:00pm - 6:30am shift"
## [143] "Sr. Data Analyst"
## [144] "Senior Scientist - Toxicologist - Product Integrity (Stewardship)"
## [145] "Senior Machine Learning Engineer"
## [146] "Data Scientist- Industrial Discrete Sector Industry"
## [147] "Senior Principal Data Scientist (Python/R)"
## [148] "Data Scientist(s)/Machine Learning Engineer"
## [149] "Scientist / Group Lead, Cancer Biology"
## [150] "Manager, Field Application Scientist, Southeast"
## [151] "COMPUTER SCIENTIST - ENGINEER - RESEARCH COMPUTER SCIENTIST - SIGNAL PROCESSING"
## [152] "Machine Learning Scientist / Engineer"
## [153] "Data Science Analyst"
## [154] "COMPUTER SCIENTIST - ENGINEER - RESEARCH COMPUTER SCIENTIST - TRANSPORTATION TECHNOLOGY"
## [155] "Software Engineer - Machine Learning & Data Science (Applied Intelligence Services Team)"
## [156] "Clinical Data Analyst"
## [157] "Data Scientist Technical Specialist"
## [158] "Data Science Manager"
## [159] "Big Data Engineer"
## [160] "Data Architect"
## [161] "Aviation AI/ML Data Scientist"
## [162] "Machine Learning Engineer, Sr."
## [163] "Information Systems Engineering Specialist (Engineering Scientist)"
## [164] "Scientist/Research Associate-Metabolic Engineering"
## [165] "Vice President, Biometrics and Clinical Data Management"
## [166] "Enterprise Data Analyst (Enterprise Portfolio Management Office)"
```

```
## [167] "Lead Data Scientist - Network Analysis and Control"
## [168] "Sr. Research Associate/ Scientist, NGS prep & Molecular Genomics"
## [169] "Developer III - Data Science"
## [170] "Hydrogen/Tritium Materials Scientist (Experienced)"
## [171] "Data Scientist/Data Analytics Practitioner"
## [172] "AI/ML - Machine Learning Scientist, Siri Understanding"
##
## $`Salary Estimate`
##  [1] "$137K-$171K (Glassdoor est.)" "$75K-$131K (Glassdoor est.)"
##  [3] "$79K-$131K (Glassdoor est.)"  "$99K-$132K (Glassdoor est.)"
##  [5] "$90K-$109K (Glassdoor est.)"  "$101K-$165K (Glassdoor est.)"
##  [7] "$56K-$97K (Glassdoor est.)"   "$79K-$106K (Glassdoor est.)"
##  [9] "$71K-$123K (Glassdoor est.)"  "$90K-$124K (Glassdoor est.)"
## [11] "$91K-$150K (Glassdoor est.)"  "$141K-$225K (Glassdoor est.)"
## [13] "$145K-$225K(Employer est.)"   "$79K-$147K (Glassdoor est.)"
## [15] "$122K-$146K (Glassdoor est.)" "$112K-$116K (Glassdoor est.)"
## [17] "$110K-$163K (Glassdoor est.)" "$124K-$198K (Glassdoor est.)"
## [19] "$79K-$133K (Glassdoor est.)"  "$69K-$116K (Glassdoor est.)"
## [21] "$31K-$56K (Glassdoor est.)"   "$95K-$119K (Glassdoor est.)"
## [23] "$212K-$331K (Glassdoor est.)" "$66K-$112K (Glassdoor est.)"
## [25] "$128K-$201K (Glassdoor est.)" "$138K-$158K (Glassdoor est.)"
## [27] "$80K-$132K (Glassdoor est.)"  "$87K-$141K (Glassdoor est.)"
## [29] "$92K-$155K (Glassdoor est.)"  "$105K-$167K (Glassdoor est.)"
##
## $Rating
##  [1]  3.1  4.2  3.8  3.5  2.9  3.9  4.4  3.6  4.5  4.7  3.7  3.4  4.1  3.2  4.3
## [16]  2.8  5.0  4.8  3.3  2.7  2.2  2.6  4.0  2.5  4.9  2.4 -1.0  2.3  4.6  3.0
## [31]  2.1  2.0
##
## $`Company Name`
##   [1] "Healthfirst\r\n3.1"
##   [2] "ManTech\r\n4.2"
##   [3] "Analysis Group\r\n3.8"
##   [4] "INFICON\r\n3.5"
##   [5] "Affinity Solutions\r\n2.9"
##   [6] "HG Insights\r\n4.2"
##   [7] "Novartis\r\n3.9"
##   [8] "iRobot\r\n3.5"
##   [9] "Intuit - Data\r\n4.4"
##  [10] "XSELL Technologies\r\n3.6"
##  [11] "Novetta\r\n4.5"
##  [12] "1904labs\r\n4.7"
##  [13] "PNNL\r\n3.7"
##  [14] "Old World Industries\r\n3.1"
##  [15] "Mathematica Policy Research\r\n3.4"
##  [16] "Guzman & Griffin Technologies (GGTI)\r\n4.4"
##  [17] "Upside Business Travel\r\n4.1"
##  [18] "Buckman\r\n3.5"
##  [19] "Insight Enterprises, Inc.\r\n4.2"
##  [20] "Tower Health\r\n3.5"
##  [21] "Triplebyte\r\n3.2"
##  [22] "PulsePoint\r\n4.3"
##  [23] "Exponent\r\n3.5"
##  [24] "Guardian Life\r\n3.5"
```

```
##  [25] "Spectrum Communications and Consulting\r\n3.4"
##  [26] "Oversight Systems\r\n4.7"
##  [27] "LSQ\r\n4.2"
##  [28] "MIT Lincoln Laboratory\r\n3.8"
##  [29] "Kingfisher Systems\r\n4.5"
##  [30] "Formation\r\n2.8"
##  [31] "Cohere Health\r\n5.0"
##  [32] "Acuity Insurance\r\n4.8"
##  [33] "Chef\r\n3.6"
##  [34] "Puget Sound Energy\r\n3.3"
##  [35] "Sandhills Global\r\n2.7"
##  [36] "A Place for Mom\r\n2.7"
##  [37] "Great-Circle Technologies\r\n2.2"
##  [38] "Edmunds.com\r\n3.4"
##  [39] "Cambridge Associates, LLC\r\n3.1"
##  [40] "Liberty Mutual Insurance\r\n3.4"
##  [41] "Cenlar\r\n2.6"
##  [42] "Arsenal Biosciences\r\n5.0"
##  [43] "Eversight\r\n4.2"
##  [44] "Pfizer\r\n4.1"
##  [45] "Klaviyo\r\n4.8"
##  [46] "Intellectual Ventures\r\n3.3"
##  [47] "GovTech\r\n3.7"
##  [48] "Quick Base\r\n4.3"
##  [49] "Giving Assistant\r\n4.8"
##  [50] "Takeda\r\n3.7"
##  [51] "Netskope\r\n4.0"
##  [52] "IT Concepts\r\n4.8"
##  [53] "iSeatz\r\n3.5"
##  [54] "Summa Health System\r\n3.7"
##  [55] "Benson Hill\r\n3.5"
##  [56] "Twitter\r\n4.1"
##  [57] "Postmates - Corporate HQ\r\n3.2"
##  [58] "Envision LLC\r\n4.5"
##  [59] "Swiss Re\r\n3.8"
##  [60] "Systems & Technology Research\r\n4.5"
##  [61] "Dermalogica\r\n3.8"
##  [62] "Bayview Asset Management\r\n3.7"
##  [63] "Via Transportation\r\n3.7"
##  [64] "Grid Dynamics\r\n4.0"
##  [65] "Tempus Labs\r\n3.3"
##  [66] "CareDx\r\n2.5"
##  [67] "IZEA\r\n4.2"
##  [68] "Autodesk\r\n4.0"
##  [69] "Caterpillar\r\n3.7"
##  [70] "New England Biolabs\r\n4.9"
##  [71] "Allied Solutions\r\n3.4"
##  [72] "The Knot Worldwide\r\n3.5"
##  [73] "IFG Companies\r\n2.9"
##  [74] "Amyris\r\n3.3"
##  [75] "AstraZeneca\r\n4.0"
##  [76] "Powertek\r\n3.6"
##  [77] "Object Partners\r\n4.7"
##  [78] "The Mom Project\r\n4.9"
```

```
##  [79] "Lightspeed Systems\r\n4.3"
##  [80] "Stripe\r\n4.0"
##  [81] "Comprehensive Healthcare\r\n2.6"
##  [82] "Fullpower Technologies, Inc.\r\n4.5"
##  [83] "Mars\r\n3.9"
##  [84] "NuWave Solutions\r\n4.4"
##  [85] "Merrick Bank\r\n3.6"
##  [86] "QOMPLX\r\n3.5"
##  [87] "GutCheck\r\n3.8"
##  [88] "Inter-American Development Bank\r\n3.5"
##  [89] "Avlino\r\n4.9"
##  [90] "Stratagem Group\r\n4.4"
##  [91] "Evidation\r\n4.1"
##  [92] "Tecolote Research\r\n3.8"
##  [93] "Tivity Health\r\n3.2"
##  [94] "hc1\r\n2.9"
##  [95] "HP Inc.\r\n4.1"
##  [96] "SAIC\r\n3.7"
##  [97] "AllianceBernstein\r\n3.2"
##  [98] "Big Huge Games\r\n4.9"
##  [99] "Maxar Technologies\r\n3.5"
## [100] "Phantom AI\r\n5.0"
## [101] "Noblis\r\n4.0"
## [102] "Spring Health\r\n3.6"
## [103] "ClearEdge\r\n4.0"
## [104] "GetWellNetwork\r\n4.8"
## [105] "TACG Solutions\r\n4.5"
## [106] "Scoop\r\n4.7"
## [107] "Montway Inc\r\n3.4"
## [108] "Juniper Networks\r\n3.8"
## [109] "Notion Labs\r\n5.0"
## [110] "Lendio\r\n4.9"
## [111] "Direct Agents\r\n4.4"
## [112] "NAVEX Global\r\n3.3"
## [113] "Upstart\r\n4.2"
## [114] "AppLovin\r\n4.8"
## [115] "ISO New England\r\n3.8"
## [116] "Relativity\r\n3.7"
## [117] "Tempo Automation\r\n3.3"
## [118] "MITRE\r\n3.3"
## [119] "Expedition Technology, Inc.\r\n5.0"
## [120] "Evidera\r\n3.8"
## [121] "Plymouth Rock Assurance\r\n3.4"
## [122] "Crown Bioscience\r\n2.4"
## [123] "GNS Healthcare\r\n2.9"
## [124] "OneMagnify\r\n4.4"
## [125] "SPECTRUM\r\n2.9"
## [126] "Advanced BioScience Laboratories\r\n2.7"
## [127] "Procore Technologies\r\n4.2"
## [128] "Ritedose\r\n3.5"
## [129] "Covid-19 Search Partners"
## [130] "bioMérieux\r\n4.2"
## [131] "Radical Convergence"
## [132] "Leidos\r\n3.5"
```

```
## [133] "Demandbase\r\n4.5"
## [134] "Shelter Insurance\r\n4.1"
## [135] "USAC\r\n2.7"
## [136] "General Dynamics Information Technology\r\n3.4"
## [137] "Offerpad\r\n4.4"
## [138] "Magna International Inc.\r\n3.5"
## [139] "United BioSource\r\n2.3"
## [140] "Kelly\r\n3.4"
## [141] "C3.ai\r\n4.7"
## [142] "Quartet Health\r\n3.9"
## [143] "Midland Credit Management\r\n3.3"
## [144] "Resurgent Capital Services\r\n4.4"
## [145] "webfx.com\r\n4.7"
## [146] "Argo Group US\r\n3.5"
## [147] "BWX Technologies\r\n3.3"
## [148] "Life360\r\n3.9"
## [149] "MassMutual\r\n3.7"
## [150] "Natera\r\n3.9"
## [151] "Genentech\r\n4.0"
## [152] "Ntrepid\r\n4.2"
## [153] "Constant Contact\r\n3.6"
## [154] "Sage Intacct\r\n4.7"
## [155] "Shape Security\r\n4.1"
## [156] "SkillSoniq\r\n5.0"
## [157] "Joby Aviation\r\n4.3"
## [158] "Cook Children's Health Care System\r\n3.8"
## [159] "Rubius Therapeutics\r\n3.8"
## [160] "GreatAmerica Financial Services\r\n4.6"
## [161] "Coverent\r\n4.1"
## [162] "Mteq\r\n3.7"
## [163] "Rocket Lawyer\r\n4.4"
## [164] "Alion Science & Technology\r\n3.6"
## [165] "Protolabs\r\n3.7"
## [166] "Quest Integrity\r\n2.9"
## [167] "Phoenix Operations Group\r\n5.0"
## [168] "Dice.com\r\n3.4"
## [169] "Southwest Research Institute\r\n3.9"
## [170] "The Buffalo Group\r\n4.3"
## [171] "Central California Alliance for Health\r\n3.5"
## [172] "Security Finance Corporation of Spartanburg\r\n3.1"
## [173] "Opendoor\r\n3.6"
## [174] "Global Data Management Inc\r\n4.5"
## [175] "Photon Infotech\r\n3.0"
## [176] "REE\r\n5.0"
## [177] "Riverside Research Institute\r\n3.6"
## [178] "T. Rowe Price\r\n3.6"
## [179] "Encode, Inc."
## [180] "Brighthouse Financial\r\n3.8"
## [181] "II-VI Incorporated\r\n3.3"
## [182] "Surya Systems\r\n4.6"
## [183] "PayPal\r\n3.8"
## [184] "Predictive Research Inc\r\n3.9"
## [185] "1010data\r\n3.1"
## [186] "Gigya\r\n3.6"
```

```
## [187] "Genesis Research\r\n5.0"
## [188] "Sanofi\r\n3.7"
## [189] "XPO Logistics\r\n3.7"
## [190] "Trace Data\r\n3.9"
## [191] "Descript\r\n4.3"
## [192] "Rincon Research Corporation\r\n4.2"
## [193] "Better Hire\r\n4.0"
## [194] "Parker Hannifin\r\n3.3"
## [195] "Gallup\r\n4.1"
## [196] "Insider Inc\r\n3.3"
## [197] "Rapid Value Solutions\r\n3.9"
## [198] "Battelle\r\n3.1"
## [199] "The Drive Media, Inc.\r\n5.0"
## [200] "Pacific Northwest National Laboratory\r\n3.7"
## [201] "US Pharmacopeia\r\n3.2"
## [202] "Itlize Global\r\n4.6"
## [203] "eBay\r\n3.5"
## [204] "Paige\r\n5.0"
## [205] "ABIOMED\r\n4.1"
## [206] "Comcast\r\n3.5"
## [207] "Metronome, LLC\r\n3.2"
## [208] "Lawrence Livermore National Lab\r\n4.7"
## [209] "FHLBank Pittsburgh\r\n3.8"
## [210] "Jacobs\r\n3.6"
## [211] "Underwriters Laboratories\r\n3.3"
## [212] "Altus Group\r\n3.7"
## [213] "Jobot\r\n5.0"
## [214] "Trovetechs Inc"
## [215] "Oshkosh Corporation\r\n4.2"
## [216] "Mackin\r\n3.4"
## [217] "PETADATA"
## [218] "VBeyond Corporation\r\n4.4"
## [219] "Take-Two\r\n3.7"
## [220] "Colony Brands\r\n3.7"
## [221] "Capio Group\r\n4.1"
## [222] "SleePare\r\n3.4"
## [223] "ShorePoint\r\n4.5"
## [224] "Dolphin\r\n3.5"
## [225] "TE Connectivity\r\n3.6"
## [226] "State of Virginia\r\n3.2"
## [227] "TA Digital\r\n3.7"
## [228] "Market America Inc\r\n4.0"
## [229] "TrueAccord\r\n3.4"
## [230] "ALTA IT Services\r\n3.9"
## [231] "Kollasoft Inc.\r\n3.2"
## [232] "ASRC Federal Holding Company\r\n3.4"
## [233] "Adwait Algorithm\r\n4.4"
## [234] "Cambridge FX\r\n3.5"
## [235] "Metromile\r\n3.8"
## [236] "Criteo\r\n3.9"
## [237] "Advance Sourcing Concepts\r\n3.4"
## [238] "Enterprise Solutions Inc\r\n3.8"
## [239] "Microagility"
## [240] "Conch Technologies, Inc\r\n4.6"
```

```
## [241] "GSK\r\n3.9"
## [242] "Rainmaker Resources, LLC"
## [243] "22nd Century Technologies\r\n3.7"
## [244] "Huxley\r\n3.3"
## [245] "FM Systems\r\n3.4"
## [246] "B4Corp"
## [247] "Blue Cross and Blue Shield of North Carolina\r\n3.7"
## [248] "Jane Street\r\n4.8"
## [249] "SSATI\r\n5.0"
## [250] "Solving IT International Inc\r\n3.4"
## [251] "The Davey Tree Expert Company\r\n3.3"
## [252] "Centauri\r\n4.6"
## [253] "Stride Search"
## [254] "Software Engineering Institute\r\n2.6"
## [255] "TechProjects\r\n4.8"
## [256] "7Park Data\r\n3.9"
## [257] "Ameritas Life Insurance Corp\r\n3.0"
## [258] "Western Digital\r\n3.5"
## [259] "Shimento, Inc.\r\n2.9"
## [260] "Averity\r\n5.0"
## [261] "Praxis Engineering\r\n4.7"
## [262] "Point72 Ventures"
## [263] "Johns Hopkins University Applied Physics Laboratory\r\n4.5"
## [264] "Cambridge Mobile Telematics\r\n4.9"
## [265] "Blend360\r\n4.6"
## [266] "Nolij Consulting\r\n3.9"
## [267] "Hatch Data Inc"
## [268] "Compass Consulting Group\r\n4.7"
## [269] "SolutionIT, Inc.\r\n4.4"
## [270] "Perspecta\r\n3.2"
## [271] "Smith Hanley Associates\r\n4.5"
## [272] "Allen Institute\r\n3.5"
## [273] "Eliassen Group\r\n4.4"
## [274] "Bayside Solutions\r\n3.1"
## [275] "Evolve Vacation Rental\r\n3.5"
## [276] "AgreeYa Solutions\r\n3.8"
## [277] "Carolina Power & Light Co\r\n3.7"
## [278] "New Iron Group, Inc.\r\n5.0"
## [279] "Travelers\r\n4.0"
## [280] "Twitch\r\n3.6"
## [281] "Biogen\r\n3.6"
## [282] "HireAi"
## [283] "Mentor Graphics\r\n4.1"
## [284] "WCG (WIRB-Copernicus Group)\r\n3.6"
## [285] "Visionary Integration Professionals\r\n4.3"
## [286] "Dynetics\r\n4.0"
## [287] "Navy Federal Credit Union\r\n3.9"
## [288] "Exact Sciences Corporation\r\n4.0"
## [289] "Community Behavioral Health\r\n3.6"
## [290] "Reynolds American\r\n3.3"
## [291] "LifeOmic\r\n5.0"
## [292] "Visionist, Inc.\r\n4.9"
## [293] "Navio"
## [294] "Concerto HealthAI\r\n3.3"
```

```
## [295] "Evolvinc"
## [296] "PROPRIUS\r\n5.0"
## [297] "TECHNOCRAFT Solutions\r\n3.4"
## [298] "Latitude, Inc.\r\n4.1"
## [299] "Royce Geospatial\r\n5.0"
## [300] "CyberCoders\r\n4.2"
## [301] "Booz Allen Hamilton Inc.\r\n3.7"
## [302] "Burns & McDonnell\r\n3.8"
## [303] "InvenTech Info\r\n4.8"
## [304] "Robert Half\r\n3.5"
## [305] "Conflux Systems Inc.\r\n4.5"
## [306] "Voice\r\n3.4"
## [307] "Falcon IT & Staffing Solutions"
## [308] "DataLab USA\r\n3.6"
## [309] "Werner Enterprises Inc\r\n3.1"
## [310] "PeopleCom\r\n5.0"
## [311] "VANTA Partners\r\n5.0"
## [312] "Blue Icy Water, LLC"
## [313] "Farmer's Business Network, Inc.\r\n3.5"
## [314] "Sonde Health"
## [315] "Maxiom\r\n5.0"
## [316] "Change Healthcare\r\n2.7"
## [317] "DCS Corp\r\n4.1"
## [318] "Hive (CA)\r\n2.1"
## [319] "Hackensack Meridian Health\r\n3.3"
## [320] "Net2Source Inc.\r\n3.2"
## [321] "The Trade Desk\r\n3.2"
## [322] "IBM\r\n3.7"
## [323] "Knowesis Inc.\r\n4.4"
## [324] "MoTek Technologies\r\n3.1"
## [325] "HPOne\r\n3.5"
## [326] "Blue Cloak LLC"
## [327] "TBWA\\Chiat\\Day\r\n2.7"
## [328] "ThreeBridge Solutions\r\n3.5"
## [329] "Numeric, LLC\r\n3.2"
## [330] "Centraprise\r\n4.2"
## [331] "DW Simpson\r\n4.2"
## [332] "LinQuest\r\n3.9"
## [333] "Trexquant Investment\r\n4.0"
## [334] "Fleetcor\r\n3.7"
## [335] "Radiant Digital\r\n4.5"
## [336] "Child Care Aware of America\r\n2.8"
## [337] "IntelliPro Group Inc.\r\n4.1"
## [338] "USI\r\n3.4"
## [339] "Apex Systems\r\n3.9"
## [340] "Pragmatics, Inc.\r\n2.9"
## [341] "Crossover Health\r\n3.5"
## [342] "Lorven Technologies Inc\r\n4.0"
## [343] "Gap Inc.\r\n3.5"
## [344] "Tygart Technology, Inc\r\n4.7"
## [345] "Murray Resources\r\n4.6"
## [346] "New York Technology Partners\r\n4.0"
## [347] "Two95 International Inc.\r\n4.0"
## [348] "Sophinea"
```

```
## [349] "CRS Group\r\n4.7"
## [350] "Blackstone Talent Group\r\n3.5"
## [351] "Roche\r\n4.1"
## [352] "Creative Circle\r\n3.6"
## [353] "Blue Horizon Tek Solutions\r\n5.0"
## [354] "Sharpedge Solutions Inc\r\n4.7"
## [355] "Alaka`ina Foundation Family of Companies\r\n3.6"
## [356] "Hexagon US Federal\r\n2.7"
## [357] "Monte Rosa Therapeutics"
## [358] "Comtech Global Inc\r\n4.0"
## [359] "Aveshka, Inc.\r\n3.8"
## [360] "10x Genomics\r\n4.2"
## [361] "CompuForce"
## [362] "1-800-Flowers\r\n2.7"
## [363] "Aptive\r\n3.5"
## [364] "JCD Staffing\r\n5.0"
## [365] "Thumbtack\r\n3.9"
## [366] "MILVETS Systems Technology, Inc.\r\n3.4"
## [367] "Apple\r\n4.1"
## [368] "Kforce\r\n4.1"
## [369] "OppLoans\r\n4.4"
## [370] "Brillient\r\n3.9"
## [371] "Xator Corporation\r\n2.9"
## [372] "HAN IT Staffing Inc.\r\n4.6"
## [373] "Infinitive Inc\r\n3.4"
## [374] "New Relic\r\n4.7"
## [375] "ICW Group\r\n3.3"
## [376] "NYSTEC\r\n3.8"
## [377] "E3 Federal Solutions\r\n4.5"
## [378] "Peraton\r\n3.4"
## [379] "Group O\r\n3.1"
## [380] "CaptiveAire\r\n4.1"
## [381] "Temboo\r\n3.9"
## [382] "Kibo"
## [383] "AeroVironment\r\n4.2"
## [384] "Applied Research Laboratories\r\n3.8"
## [385] "Conagen\r\n2.0"
## [386] "Alector\r\n4.8"
## [387] "Homology Medicines, Inc.\r\n4.4"
## [388] "Inland Empire Health Plan\r\n3.3"
## [389] "Sandia National Laboratories\r\n3.8"
## [390] "CIA\r\n3.8"
## [391] "Maven Wave Partners\r\n4.5"
## [392] "Ovative Group\r\n4.3"
## [393] "Kognetics\r\n3.6"
## [394] "Envision Healthcare\r\n2.9"
## [395] "ConsumerTrack\r\n3.2"
## [396] "Meridian Knowledge Solutions\r\n4.4"
## [397] "UST Global\r\n4.2"
## [398] "IMG Systems\r\n3.2"
## [399] "Trident Systems Inc\r\n3.4"
## [400] "GrainBridge, LLC"
## [401] "First Health Group\r\n3.2"
## [402] "Sprezzatura Management Consulting"
```

```
## [403] "Progress Rail, A Caterpillar Company\r\n2.8"
## [404] "Axiologic Solutions\r\n4.5"
## [405] "Indigo Slate\r\n3.0"
## [406] "Cubic\r\n3.3"
## [407] "Advanced Bio-Logic Solutions Corp\r\n4.0"
## [408] "Alignment Healthcare\r\n3.5"
## [409] "WGSN\r\n3.5"
## [410] "ISYS Technologies, Inc.\r\n3.6"
## [411] "TransVoyant\r\n3.0"
## [412] "Geotab\r\n4.3"
## [413] "EGlobalTech\r\n3.7"
## [414] "Central Business Solutions, Inc\r\n3.0"
## [415] "KeHE Distributors\r\n2.5"
## [416] "Moxie Software\r\n3.0"
## [417] "Unicom Technologies INC\r\n4.7"
## [418] "Americo Life\r\n3.3"
## [419] "Tokio Marine HCC\r\n3.3"
## [420] "CACI International\r\n3.5"
## [421] "Berico Technologies"
## [422] "Kehe Food Distributors"
## [423] "Pactera Edge"
## [424] "Qurate Retail Group\r\n3.6"
## [425] "A-Line Staffing Solutions\r\n4.1"
## [426] "Clear Ridge Defense"
## [427] "Criterion Systems, Inc.\r\n3.8"
## [428] "Foundation Medicine\r\n4.0"
## [429] "TRANZACT\r\n3.6"
## [430] "JKGT"
## [431] "AccessHope"
## [432] "ChaTeck Incorporated\r\n5.0"
##
## $Location
##   [1] "New York, NY"              "Chantilly, VA"
##   [3] "Boston, MA"                "Newton, MA"
##   [5] "Santa Barbara, CA"         "Cambridge, MA"
##   [7] "Bedford, MA"               "San Diego, CA"
##   [9] "Chicago, IL"               "Herndon, VA"
##  [11] "Saint Louis, MO"           "Richland, WA"
##  [13] "Northbrook, IL"            "Washington, DC"
##  [15] "Remote"                    "Memphis, TN"
##  [17] "Plano, TX"                 "West Grove, PA"
##  [19] "Phoenix, AZ"               "Appleton, WI"
##  [21] "Atlanta, GA"               "Orlando, FL"
##  [23] "Lexington, MA"             "McLean, VA"
##  [25] "San Francisco, CA"         "Sheboygan, WI"
##  [27] "United States"             "Bothell, WA"
##  [29] "Lincoln, NE"               "Overland Park, KS"
##  [31] "Santa Monica, CA"          "Portsmouth, NH"
##  [33] "Ewing, NJ"                 "South San Francisco, CA"
##  [35] "Palo Alto, CA"             "Bellevue, WA"
##  [37] "New Orleans, LA"           "Akron, OH"
##  [39] "Fort Wayne, IN"            "Woburn, MA"
##  [41] "Carson, CA"                "Coral Gables, FL"
##  [43] "Santa Clara, CA"           "Brisbane, CA"
```

```
##  [45] "Winter Park, FL"           "Redwood City, CA"
##  [47] "Peoria, IL"                "Ipswich, MA"
##  [49] "Carmel, IN"                "Emeryville, CA"
##  [51] "Gaithersburg, MD"          "Longmont, CO"
##  [53] "Austin, TX"                "Yakima, WA"
##  [55] "Santa Cruz, CA"            "Springfield, VA"
##  [57] "Alexandria, VA"            "Utah"
##  [59] "Reston, VA"                "Denver, CO"
##  [61] "New Jersey"                "Aurora, CO"
##  [63] "Hill AFB, UT"              "Chandler, AZ"
##  [65] "Indianapolis, IN"          "Nashville, TN"
##  [67] "Timonium, MD"              "Burlingame, CA"
##  [69] "Annapolis Junction, MD"    "Bethesda, MD"
##  [71] "Dayton, OH"                "Schaumburg, IL"
##  [73] "Cupertino, CA"             "Lehi, UT"
##  [75] "Culver City, CA"           "Lake Oswego, OR"
##  [77] "San Mateo, CA"             "Holyoke, MA"
##  [79] "Woodbridge, NJ"            "Dearborn, MI"
##  [81] "Maryland Heights, MO"      "Rockville, MD"
##  [83] "Carpinteria, CA"           "Columbia, SC"
##  [85] "Hauppauge, NY"             "Fort Meade, MD"
##  [87] "Columbia, MO"              "Vicksburg, MS"
##  [89] "Birmingham, AL"            "Blue Bell, PA"
##  [91] "Cincinnati, OH"            "Harrisburg, PA"
##  [93] "Oak Ridge, TN"             "San Carlos, CA"
##  [95] "Waltham, MA"               "Fort Worth, TX"
##  [97] "Smithfield, RI"            "Cedar Rapids, IA"
##  [99] "Fort Belvoir, VA"          "Linthicum Heights, MD"
## [101] "Maple Plain, MN"           "Tulsa, OK"
## [103] "Baltimore, MD"             "Oklahoma City, OK"
## [105] "Scotts Valley, CA"         "Spartanburg, SC"
## [107] "Hartford, CT"              "Beavercreek, OH"
## [109] "Norfolk, VA"               "Charlotte, NC"
## [111] "Champaign, IL"             "Texas"
## [113] "Hoboken, NJ"               "Lebanon, IN"
## [115] "Oakland, CA"               "Melbourne, FL"
## [117] "Cleveland, OH"             "Norwell, MA"
## [119] "San Jose, CA"              "Piscataway, NJ"
## [121] "Danvers, MA"               "Vienna, VA"
## [123] "Livermore, CA"             "Pittsburgh, PA"
## [125] "Irvine, CA"                "Oshkosh, WI"
## [127] "Menlo Park, CA"            "Dallas, TX"
## [129] "Arlington, VA"             "Monroe, WI"
## [131] "Sacramento, CA"            "Hampton, VA"
## [133] "Richmond, VA"              "Monterey, CA"
## [135] "Woodlawn, MD"              "Ann Arbor, MI"
## [137] "Concord, CA"               "Durham, NC"
## [139] "Kent, OH"                  "Laurel, MD"
## [141] "Columbia, MD"              "Falls Church, VA"
## [143] "Thousand Oaks, CA"         "Edison, NJ"
## [145] "Adelphi, MD"               "Seattle, WA"
## [147] "Sunnyvale, CA"             "Fremont, CA"
## [149] "Hamilton, NJ"              "Huntsville, AL"
## [151] "Merrifield, VA"            "Madison, WI"
```

```
## [153] "Philadelphia, PA"           "Winston-Salem, NC"
## [155] "Raleigh, NC"                "Burbank, CA"
## [157] "San Ramon, CA"              "Oxnard, CA"
## [159] "Kansas City, MO"            "Jersey City, NJ"
## [161] "Manchester, NH"             "Winters, TX"
## [163] "Brooklyn, NY"               "Germantown, MD"
## [165] "Omaha, NE"                  "Open Fork, VA"
## [167] "Ashburn, VA"                "Lombard, IL"
## [169] "Alpharetta, GA"             "Boulder, CO"
## [171] "Mountain View, CA"          "Trumbull, CT"
## [173] "Sterling, VA"               "Foster City, CA"
## [175] "Frederick, MD"              "Colorado Springs, CO"
## [177] "Southfield, MI"             "San Clemente, CA"
## [179] "The Woodlands, TX"          "Pleasanton, CA"
## [181] "Wilmington, DE"             "Fort Sam Houston, TX"
## [183] "Lexington Park, MD"         "Patuxent, Anne Arundel, MD"
## [185] "Fairfax, VA"                "San Antonio, TX"
## [187] "Silver Spring, MD"          "Portland, OR"
## [189] "Simi Valley, CA"            "New Bedford, MA"
## [191] "Rancho Cucamonga, CA"       "Collegeville, PA"
## [193] "Minneapolis, MN"            "Gahanna, OH"
## [195] "California"                 "Wellesley, MA"
## [197] "Washington, VA"             "Orange, CA"
## [199] "Bridgeport, WV"             "Oakville, CA"
## [201] "Naperville, IL"             "Houston, TX"
## [203] "Redmond, WA"                "West Chester, PA"
## [205] "Quantico, VA"               "Fort Lee, NJ"
## [207] "Irwindale, CA"
##
## $Headquarters
##   [1] "New York, NY"               "Herndon, VA"
##   [3] "Boston, MA"                 "Bad Ragaz, Switzerland"
##   [5] "Santa Barbara, CA"          "Basel, Switzerland"
##   [7] "Bedford, MA"                "Mountain View, CA"
##   [9] "Chicago, IL"                "Mc Lean, VA"
##  [11] "Saint Louis, MO"            "Richland, WA"
##  [13] "Northbrook, IL"             "Princeton, NJ"
##  [15] "Mays Landing, NJ"           "Washington, DC"
##  [17] "Memphis, TN"                "Tempe, AZ"
##  [19] "Reading, PA"                "San Francisco, CA"
##  [21] "Menlo Park, CA"             "Atlanta, GA"
##  [23] "Orlando, FL"                "Lexington, MA"
##  [25] "Falls Church, VA"           "Sheboygan, WI"
##  [27] "Seattle, WA"                "Bellevue, WA"
##  [29] "Lincoln, NE"                "Chantilly, VA"
##  [31] "Santa Monica, CA"           "Ewing, NJ"
##  [33] "South San Francisco, CA"    "Palo Alto, CA"
##  [35] "Singapore, Singapore"       "Cambridge, MA"
##  [37] "OSAKA, Japan"               "Santa Clara, CA"
##  [39] "Vienna, VA"                 "New Orleans, LA"
##  [41] "Akron, OH"                  "Zurich, Switzerland"
##  [43] "Woburn, MA"                 "Carson, CA"
##  [45] "Coral Gables, FL"           "San Ramon, CA"
##  [47] "Brisbane, CA"               "Winter Park, FL"
```

```
##  [49] "San Rafael, CA"            "Deerfield, IL"
##  [51] "Ipswich, MA"               "Carmel, IN"
##  [53] "Chevy Chase, MD"           "Hartford, CT"
##  [55] "Emeryville, CA"            "Cambridge, United Kingdom"
##  [57] "Rockville, MD"             "Minneapolis, MN"
##  [59] "Austin, TX"                "Yakima, WA"
##  [61] "Santa Cruz, CA"            "South Jordan, UT"
##  [63] "Reston, VA"                "Denver, CO"
##  [65] "Holmdel, NJ"               "Aurora, CO"
##  [67] "San Mateo, CA"             "Goleta, CA"
##  [69] "Franklin, TN"              "Indianapolis, IN"
##  [71] "Lutherville Timonium, MD"  "Westminster, CO"
##  [73] "Burlingame, CA"            "Annapolis Junction, MD"
##  [75] "Bethesda, MD"              "Beavercreek, OH"
##  [77] "Schaumburg, IL"            "Sunnyvale, CA"
##  [79] "Lehi, UT"                  "Lake Oswego, OR"
##  [81] "Holyoke, MA"               "Dulles, VA"
##  [83] "San Diego, CA"             "Detroit, MI"
##  [85] "Stamford, CT"              "Carpinteria, CA"
##  [87] "Columbia, SC"              "-1"
##  [89] "Marcy-l'Etoile, France"    "Columbia, MO"
##  [91] "Fairfax, VA"               "Chandler, AZ"
##  [93] "Aurora, Canada"            "Blue Bell, PA"
##  [95] "Troy, MI"                  "Redwood City, CA"
##  [97] "Greenville, SC"            "Arlington, VA"
##  [99] "Harrisburg, PA"            "Hamilton, Bermuda"
## [101] "Lynchburg, VA"             "Springfield, MA"
## [103] "San Carlos, CA"            "Waltham, MA"
## [105] "San Jose, CA"              "Jersey City, NJ"
## [107] "Fort Worth, TX"            "Cedar Rapids, IA"
## [109] "McLean, VA"                "Lorton, VA"
## [111] "Maple Plain, MN"           "Kent, WA"
## [113] "Woodbine, MD"              "San Antonio, TX"
## [115] "Scotts Valley, CA"         "Spartanburg, SC"
## [117] "Woodbridge, NJ"            "Chennai, India"
## [119] "Tel Aviv-Yafo, Israel"     "Baltimore, MD"
## [121] "Manalapan, NJ"             "Charlotte, NC"
## [123] "Saxonburg, PA"             "Bristol, PA"
## [125] "Bangalore, India"          "Hoboken, NJ"
## [127] "Paris, France"             "Greenwich, CT"
## [129] "Santa Ana, CA"             "Houston, TX"
## [131] "Tucson, AZ"                "Birmingham, AL"
## [133] "Cleveland, OH"             "Pleasanton, CA"
## [135] "Columbus, OH"              "Piscataway, NJ"
## [137] "Danvers, MA"               "Philadelphia, PA"
## [139] "Livermore, CA"             "Pittsburgh, PA"
## [141] "Dallas, TX"                "Toronto, Canada"
## [143] "Irvine, CA"                "Hillsborough, NJ"
## [145] "Oshkosh, WI"               "Fremont, CA"
## [147] "Monroe, WI"                "Goteborg, Sweden"
## [149] "Lake Buena Vista, FL"      "Schaffhausen, Switzerland"
## [151] "Richmond, VA"              "Newark, CA"
## [153] "Greensboro, NC"            "Scottsdale, AZ"
## [155] "Irving, TX"                "Beltsville, MD"
```

```
## [157] "Naperville, IL"            "Brentford, United Kingdom"
## [159] "Cincinnati, OH"            "Somerset, NJ"
## [161] "London, United Kingdom"    "Raleigh, NC"
## [163] "Leesburg, VA"              "Durham, NC"
## [165] "Kent, OH"                  "Westlake Village, CA"
## [167] "North Brunswick, NJ"       "Benicia, CA"
## [169] "Laurel, MD"                "Columbia, MD"
## [171] "Danville, CA"              "Wilmington, MA"
## [173] "New York, 061"             "Reading, MA"
## [175] "Folsom, CA"                "Wilsonville, OR"
## [177] "Huntsville, AL"            "Madison, WI"
## [179] "Phila, PA"                 "Winston-Salem, NC"
## [181] "Half Moon Bay, CA"         "Los Angeles, CA"
## [183] "Hilliard, OH"              "Hanover, MD"
## [185] "Kansas City, MO"           "Bengaluru, India"
## [187] "Alpharetta, GA"            "Germantown, MD"
## [189] "Omaha, NE"                 "Clifton Park, NY"
## [191] "Livonia, MI"               "Ashburn, VA"
## [193] "Nashville, TN"             "Alexandria, VA"
## [195] "Edison, NJ"                "Ventura, CA"
## [197] "Armonk, NY"                "Trumbull, CT"
## [199] "Chadds Ford, PA"           "Saint Paul, MN"
## [201] "Glen Allen, VA"            "Aliso Viejo, CA"
## [203] "Plainsboro, NJ"            "Fairmont, WV"
## [205] "Cherry Hill, NJ"           "Itasca, IL"
## [207] "Coconut Creek, FL"         "Lombard, IL"
## [209] "Honolulu, HI"              "Carle Place, NY"
## [211] "Cupertino, CA"             "Tampa, FL"
## [213] "Totowa, NJ"                "Rome, NY"
## [215] "Milan, IL"                 "Marbella, Spain"
## [217] "Simi Valley, CA"           "Rancho Cucamonga, CA"
## [219] "Albuquerque, NM"           "Langley, VA"
## [221] "Plano, TX"                 "Albertville, AL"
## [223] "Orange, CA"                "Littleton, CO"
## [225] "Oakville, Canada"          "San Bruno, CA"
## [227] "West Chester, PA"          "Utica, MI"
## [229] "Fort Lee, NJ"
##
## $Size
## [1] "1001 to 5000 employees"  "5001 to 10000 employees"
## [3] "501 to 1000 employees"   "51 to 200 employees"
## [5] "10000+ employees"        "201 to 500 employees"
## [7] "1 to 50 employees"       "-1"
## [9] "Unknown"
##
## $Founded
##   [1] 1993 1968 1981 2000 1998 2010 1996 1990 1983 2014 2012 2016 1965 1973 1986
##  [16] 1997 2015 1945 1988 2017 2011 1967 1860 1992 2003 1951 2005 2019 1925 2008
##  [31] 1999 1978 1966 1912 1958 2013 1849 1781 1926 2006 1994 1863 1995   -1 1982
##  [46] 1974 2001 1985 1913 1971 1911 2009 1959 2007 1939 2002 1961 1963 1969 1946
##  [61] 1957 1953 1948 1850 1851 2004 1976 1918 1954 1947 1955 2018 1937 1917 1935
##  [76] 1929 1820 1952 1932 1894 1960 1788 1830 1984 1933 1880 1887 1970 1942 1980
##  [91] 1989 1908 1853 1875 1914 1898 1956 1977 1987 1896 1972 1949 1962
##
```

```
## $`Type of ownership`
##  [1] "Nonprofit Organization"         "Company - Public"
##  [3] "Private Practice / Firm"        "Company - Private"
##  [5] "Government"                     "Subsidiary or Business Segment"
##  [7] "Other Organization"             "-1"
##  [9] "Unknown"                        "Hospital"
## [11] "Self-employed"                  "College / University"
## [13] "Contract"
##
## $Industry
##  [1] "Insurance Carriers"
##  [2] "Research & Development"
##  [3] "Consulting"
##  [4] "Electrical & Electronic Manufacturing"
##  [5] "Advertising & Marketing"
##  [6] "Computer Hardware & Software"
##  [7] "Biotech & Pharmaceuticals"
##  [8] "Consumer Electronics & Appliances Stores"
##  [9] "Enterprise Software & Network Solutions"
## [10] "IT Services"
## [11] "Energy"
## [12] "Chemical Manufacturing"
## [13] "Federal Agencies"
## [14] "Internet"
## [15] "Health Care Services & Hospitals"
## [16] "Investment Banking & Asset Management"
## [17] "Aerospace & Defense"
## [18] "Utilities"
## [19] "-1"
## [20] "Express Delivery Services"
## [21] "Staffing & Outsourcing"
## [22] "Insurance Agencies & Brokerages"
## [23] "Consumer Products Manufacturing"
## [24] "Industrial Manufacturing"
## [25] "Food & Beverage Manufacturing"
## [26] "Banks & Credit Unions"
## [27] "Video Games"
## [28] "Shipping"
## [29] "Telecommunications Services"
## [30] "Lending"
## [31] "Cable, Internet & Telephone Providers"
## [32] "Real Estate"
## [33] "Venture Capital & Private Equity"
## [34] "Miscellaneous Manufacturing"
## [35] "Oil & Gas Services"
## [36] "Transportation Equipment Manufacturing"
## [37] "Telecommunications Manufacturing"
## [38] "Transportation Management"
## [39] "News Outlet"
## [40] "Architectural & Engineering Services"
## [41] "Food & Beverage Stores"
## [42] "Other Retail Stores"
## [43] "Hotels, Motels, & Resorts"
## [44] "State & Regional Agencies"
```

```
## [45] "Financial Transaction Processing"
## [46] "Timber Operations"
## [47] "Colleges & Universities"
## [48] "Travel Agencies"
## [49] "Accounting"
## [50] "Logistics & Supply Chain"
## [51] "Farm Support Services"
## [52] "Social Assistance"
## [53] "Construction"
## [54] "Department, Clothing, & Shoe Stores"
## [55] "Publishing"
## [56] "Health, Beauty, & Fitness"
## [57] "Wholesale"
## [58] "Rail"
##
## $Sector
##  [1] "Insurance"                          "Business Services"
##  [3] "Manufacturing"                      "Information Technology"
##  [5] "Biotech & Pharmaceuticals"          "Retail"
##  [7] "Oil, Gas, Energy & Utilities"       "Government"
##  [9] "Health Care"                        "Finance"
## [11] "Aerospace & Defense"                "-1"
## [13] "Transportation & Logistics"         "Media"
## [15] "Telecommunications"                 "Real Estate"
## [17] "Travel & Tourism"                   "Agriculture & Forestry"
## [19] "Education"                          "Accounting & Legal"
## [21] "Non-Profit"                         "Construction, Repair & Maintenance"
## [23] "Consumer Services"
##
## $Revenue
##  [1] "Unknown / Non-Applicable"         "$1 to $2 billion (USD)"
##  [3] "$100 to $500 million (USD)"       "$10+ billion (USD)"
##  [5] "$2 to $5 billion (USD)"           "$500 million to $1 billion (USD)"
##  [7] "$5 to $10 billion (USD)"          "$10 to $25 million (USD)"
##  [9] "$25 to $50 million (USD)"         "$50 to $100 million (USD)"
## [11] "$1 to $5 million (USD)"           "$5 to $10 million (USD)"
## [13] "Less than $1 million (USD)"       "-1"
##
## $Competitors
##   [1] "EmblemHealth, UnitedHealth Group, Aetna"
##   [2] "-1"
##   [3] "MKS Instruments, Pfeiffer Vacuum, Agilent Technologies"
##   [4] "Commerce Signals, Cardlytics, Yodlee"
##   [5] "Square, PayPal, H&R Block"
##   [6] "Leidos, CACI International, Booz Allen Hamilton"
##   [7] "Slalom, Daugherty Business Solutions"
##   [8] "Oak Ridge National Laboratory, National Renewable Energy Lab, Los Alamos National Laboratory"
##   [9] "CDW, PCM, SHI International"
##  [10] "Crossix Solutions Inc., AppNexus, The Trade Desk"
##  [11] "Northwestern Mutual"
##  [12] "Puppet, Ansible, SaltStack"
##  [13] "Enlivant, Sunrise Senior Living, Brookdale Senior Living"
##  [14] "TrueCar, Cars.com, Kelley Blue Book"
##  [15] "Travelers, Allstate, State Farm"
```

```
##  [16] "Novartis, Baxter, Pfizer"
##  [17] "Skyhigh Networks, Zscaler, NortonLifeLock"
##  [18] "Facebook, Google, Pinterest"
##  [19] "DoorDash, Uber, Grubhub"
##  [20] "Munich Re, Hannover RE, SCOR"
##  [21] "IMAGE Skincare, Aveda, Kiehl's"
##  [22] "Luxoft, EPAM, Capgemini Invent"
##  [23] "Sequenom"
##  [24] "Linqia, Collective Bias"
##  [25] "John Deere, Komatsu, CNH Industrial"
##  [26] "Thermo Fisher Scientific, Enzymatics, Illumina"
##  [27] "CUNA Mutual, SWBC, Overby-Seawell"
##  [28] "Zola Registry"
##  [29] "Colony Specialty, Markel, RLI"
##  [30] "Roche, GlaxoSmithKline, Novartis"
##  [31] "Solution Design Group, Intertech (Minnesota)"
##  [32] "Braintree, Authorize.Net, PayPal"
##  [33] "Nielsen, Zappi, SurveyMonkey"
##  [34] "The World Bank, IMF"
##  [35] "Booz Allen Hamilton, CACI International"
##  [36] "Booz Allen Hamilton, SAIC, LMI"
##  [37] "Epic, CipherHealth"
##  [38] "Copper River Shared Services, Chenega Corporation, Deloitte"
##  [39] "Battelle, General Atomics, SAIC"
##  [40] "IQVIA, ICON"
##  [41] "Arbella Insurance, Safety Insurance"
##  [42] "Engagio, Bombora, Terminus"
##  [43] "SAIC, Leidos, Northrop Grumman"
##  [44] "Bosch, Lear Corporation, Faurecia"
##  [45] "Covance, ICON"
##  [46] "Adecco, ManpowerGroup, Allegis Corporation"
##  [47] "GE Digital, Palantir Technologies, Uptake"
##  [48] "PRA Group"
##  [49] "Genomic Health, 23andMe, Illumina"
##  [50] "Bromium, FireEye, Authentic8"
##  [51] "Drip, iContact, Mailchimp"
##  [52] "Children's Health, Texas Health Resources, Baylor Scott & White Health"
##  [53] "Harris, Fibertek"
##  [54] "Monster Worldwide, CareerBuilder, Craigslist"
##  [55] "Los Alamos National Laboratory, Battelle, SRI International"
##  [56] "ManTech, Booz Allen Hamilton, Leidos"
##  [57] "Lumentum Operations, Keysight Technologies, O-Net Technologies"
##  [58] "Square, Amazon, Apple"
##  [59] "Pfizer, GlaxoSmithKline"
##  [60] "DHL Supply Chain, UPS, FedEx"
##  [61] "Raytheon Technologies, General Dynamics, MIT Lincoln Laboratory"
##  [62] "Eaton, SMC Corporation, Bosch Rexroth"
##  [63] "Advisory Board, Booz Allen Hamilton, McKinsey & Company"
##  [64] "Amazon, Apple"
##  [65] "Covidien, Boston Scientific"
##  [66] "AT&T, Verizon"
##  [67] "Los Alamos National Laboratory, NASA Jet Propulsion Laboratory, Sandia National Laboratories"
##  [68] "Fluor, Bechtel, AECOM"
##  [69] "Intertek, SGS, Bureau Veritas"
```

```
##  [70] "Lockheed Martin, Caterpillar, John Deere"
##  [71] "Activision Blizzard, Electronic Arts"
##  [72] "MediaMath, Conversant, AppNexus"
##  [73] "Pfizer, AstraZeneca, Merck"
##  [74] "Archibus, iOffice, Planon"
##  [75] "ACRT Services, Bartlett Tree Experts"
##  [76] "TASC, Vencore, Booz Allen Hamilton"
##  [77] "Seagate Technology, Toshiba"
##  [78] "Raytheon Technologies, Northrop Grumman, Booz Allen Hamilton"
##  [79] "MIT Lincoln Laboratory, Lockheed Martin, Northrop Grumman"
##  [80] "Kforce, PageGroup, Robert Half"
##  [81] "TEKsystems, Kforce, Randstad US"
##  [82] "South Carolina Electric & Gas, Virginia Electric and Power"
##  [83] "Cadence Design Systems, Synopsys, Altium Limited"
##  [84] "CGI (Nevada), Accenture, Deloitte"
##  [85] "Accenture, Deloitte, PwC"
##  [86] "Bechtel Jacobs, Black & Veatch, HNTB"
##  [87] "Adecco, Manpower"
##  [88] "Acxiom, Merkle, Epsilon (North Carolina)"
##  [89] "Amazon, Accenture, Microsoft"
##  [90] "Booz Allen Hamilton, Deloitte, ERPi"
##  [91] "TEKsystems, Insight Global, Accenture"
##  [92] "H&M, Inditex, Fast Retailing"
##  [93] "Novartis, AstraZeneca, Siemens Healthineers"
##  [94] "Aquent, 24 Seven Talent"
##  [95] "Google, Microsoft, Samsung Electronics"
##  [96] "AppDynamics, Datadog, Dynatrace"
##  [97] "Liberty Mutual Insurance, EMPLOYERS, Travelers"
##  [98] "KPMG, Accenture, Deloitte"
##  [99] "General Atomics, Boeing, Northrop Grumman"
## [100] "Los Alamos National Laboratory, Lawrence Livermore National Laboratory"
## [101] "Cognizant Technology Solutions, Infosys, Wipro"
## [102] "Humana"
## [103] "Accenture, Northrop Grumman, Xerox"
## [104] "United Natural Foods, US Foods, DPI Specialty Foods"
## [105] "LivePerson, Salesforce, SAP"
## [106] "Zurich Insurance, AXA XL, Allianz"
## [107] "CSC, ManTech, SAIC"
## [108] "Genomic Health, Myriad Genetics, The Broad Institute"
```

Limpieza de datos

Puntos a tener en cuenta para la limpieza de datos

A raíz del estudio hecho en el apartado anterior, hay que tener en cuenta los siguientes puntos para la limpieza de datos:

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 tendremos que tener en cuenta también esos casos y representar a todos los valores faltantes de forma homogénea mediante `NA`
2. La columna `Job title` tiene una gran diversidad de trabajos con una mínima variación que sería interesante tratarlos como un mismo trabajo. Para ello, habrá que definir un subconjunto de trabajos a partir del cuál tratar como iguales las variantes. Ese subconjunto será los que consideramos principales

: { data scientist, data engineer, data analyst, machine learning, machine learning expert }. Así, por ejemplo, un trabajo de e-commerce data analyst o uno de RFP data analyst será tratado bajo la categoría de data analyst

3. La variable `Company name` tiene la información del rating. Habrá que eliminar esa redundancia
4. Es interesante añadir 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.
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 tendrian 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 realidad el mismo salario. 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.

Tratamiento de la variable `Job Titles`

```
clean_data <- raw_data

# Asignamos el valor NA en todas las celdas de la tabla donde aparece un -1 o un Unknown
# Esto lo podemos hacer porque entre las variables numéricas que tenemos no hay ninguna en la que en su
clean_data[clean_data==-1 | clean_data=="Unknown" | clean_data=="Unknown / Non-Applicable"] <- NA

data_jobs_titles <- "data scientist|data engineer|data analyst|machine learning"

# Tratamos el nombre de los trabajos para considerar idénticos aquellos que tienen mínimas variaciones
clean_data <- clean_data %>%
  mutate(`Job Title`= tolower(`Job Title`) %>%
           str_extract(data_jobs_titles)
         )
```

Ahora comprobemos qué tipo de trabajos se han quedado fuera:

```
out_jobs_index <- clean_data[is.na(clean_data$`Job Title`),] %>%
  select(index) %>%
  as_vector() %>%
  unname()

out_jobs <- raw_data %>%
  filter(`index` %in% out_jobs_index) %>%
  select(`Job Title`) %>%
  distinct() %>%
  datatable()
```

Consideramos que los trabajos que se quedan fuera son demasiado específicos para el análisis que queremos hacer posteriormente. Además, el número de observaciones que perderíamos si no consideraramos el estudio posterior para ninguna de estas profesiones es de 87 lo que consideramos una cifra asumible.

```
clean_data <- clean_data[!is.na(clean_data$`Job Title`),]
```

Tratamiento de la variable `Salary Estimate`

Esta variable es interesante separarla en dos: una para el rango mínimo y otra para el rango máximo. Adicionalmente, es interesante crear una nueva a partir de estas dos que sea el rango medio.

Para ello, como hay muchas observaciones diferentes de rangos distintos nos aseguramos de forma automatizada que todos los rangos están en miles:

```
all_k <- raw_data %>%
  select(`Salary Estimate`) %>%
  distinct() %>%
  map(function(x) grepl(".*K.*K",x)) %>%
  as_vector() %>%
  all()

if(all_k){
  cat("Todas las observaciones están en miles, no hay que tener ningún cuidado especial")
}else{
  cat("¡Cuidado! Existen algunas observaciones que no están en miles, hay que tratar esas observaciones
}
```

```
## Todas las observaciones están en miles, no hay que tener ningún cuidado especial
```

```
clean_data <- clean_data %>%
  separate(`Salary Estimate`,sep="-", into=c("Salary Estimate Inf","Salary Estimate Sup")) %>%
  mutate(
    `Salary Estimate Inf` = gsub("K|\\$","",`Salary Estimate Inf`),
    `Salary Estimate Sup` = gsub("K|\\$","",`Salary Estimate Sup`)
  ) %>%
  #Lo separamos de nuevo para evitar poner todas las variaciones posibles Glassdoor est., Employer Est.
  separate(`Salary Estimate Sup`, sep="\\(", into=c("Salary Estimate Sup", "drop")) %>%
  select(-drop ) %>%
  mutate(
    `Salary Estimate Inf` = as.double(`Salary Estimate Inf`),
    `Salary Estimate Sup` = as.double(`Salary Estimate Sup`),
    `Salary Estimate Med` = (`Salary Estimate Inf` + `Salary Estimate Sup`)/2
  )
```

Tratamiento de la variable `Company Name`

```
clean_data <- clean_data %>%
  mutate(`Company Name`=str_remove_all(`Company Name`,"\n.*"))
```

Tratamiento de la variable `Headquarters` y `Location`

```
clean_data <-  clean_data %>%
  mutate(`Same Location Headquarter`= Location==Headquarters)
```

Tratamiento de la variable `Size`

Esta variable será tratada de forma análoga a `Salary Estimate`.

```
clean_data <- clean_data %>%
  mutate(
    `Size Ordered`=
           case_when(
             Size=="1 to 50 employees"              ~ 1,
             Size=="51 to 200 employees"            ~ 2,
             Size=="201 to 500 employees"           ~ 3,
             Size=="501 to 1000 employees"          ~ 4,
             Size=="1001 to 5000 employees"         ~ 5,
             Size=="5001 to 10000 employees"        ~ 6,
             Size=="10000+ employees"               ~ 7,
         ),
    Size=
      case_when(
             Size=="1 to 50 employees"              ~ "1-50",
             Size=="51 to 200 employees"            ~ "51-200",
             Size=="201 to 500 employees"           ~ "201-500",
             Size=="501 to 1000 employees"          ~ "501-1000",
             Size=="1001 to 5000 employees"         ~ "1001-5000",
             Size=="5001 to 10000 employees"        ~ "5001-10000",
             Size=="10000+ employees"               ~ "10000-Inf",
         )

) %>%
  separate(Size, sep="-", into=c("Size Inf","Size Sup")) %>%
  mutate(
    `Size Inf`=as.numeric(`Size Inf`),
    `Size Sup`=as.numeric(`Size Sup`)
  )
```

Tratamiento de la variable `Revenue`

Para esta variable, aparte de sacar los valores extremos, sería adecuada tratarla como una variable ordinal

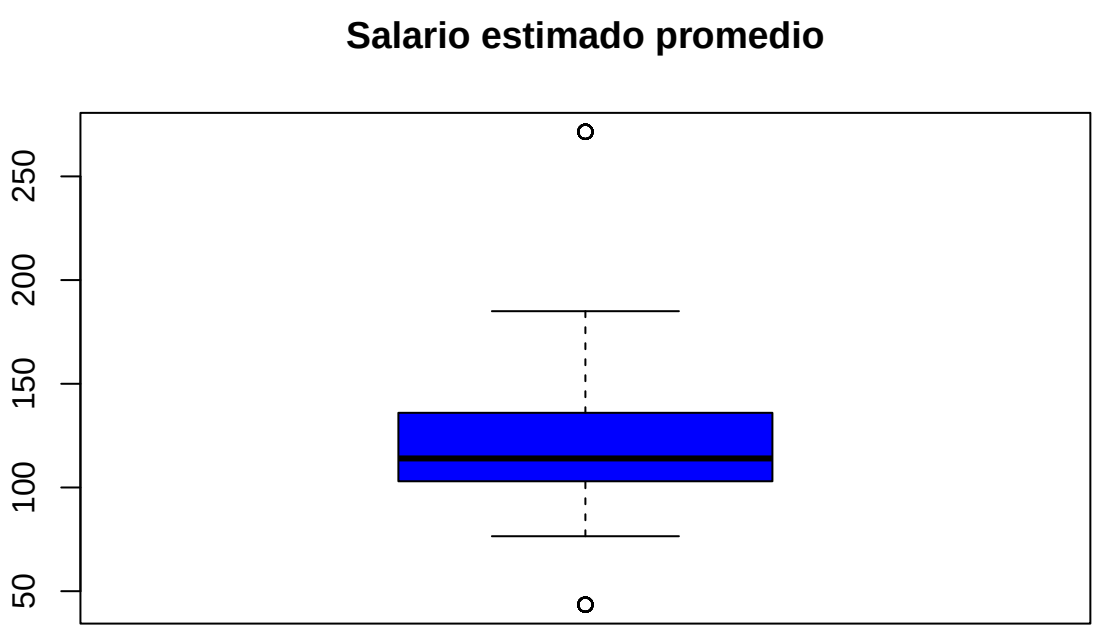
```
clean_data <- clean_data %>%
  mutate(
    `Revenue Ordered`=
           case_when(
             Revenue=="Less than $1 million (USD)"       ~ 1,
             Revenue=="$1 to $5 million (USD)"           ~ 2,
             Revenue=="$5 to $10 million (USD)"          ~ 3,
             Revenue=="$10 to $25 million (USD)"         ~ 4,
             Revenue=="$25 to $50 million (USD)"         ~ 5,
             Revenue=="$50 to $100 million (USD)"        ~ 6,
             Revenue=="$100 to $500 million (USD)"       ~ 7,
             Revenue=="$500 million to $1 billion (USD)" ~ 8,
             Revenue=="$1 to $2 billion (USD)"           ~ 9,
             Revenue=="$2 to $5 billion (USD)"           ~ 10,
             Revenue=="$5 to $10 billion (USD)"          ~ 11,
             Revenue=="$10+ billion (USD)"               ~ 12
         ),
    Revenue=
           case_when(
             Revenue=="Less than $1 million (USD)"       ~ "0-1000000",
```

```
            Revenue=="$1 to $5 million (USD)"           ~ "1000000-5000000",
            Revenue=="$5 to $10 million (USD)"          ~ "5000000-10000000",
            Revenue=="$10 to $25 million (USD)"         ~ "10000000-25000000",
            Revenue=="$25 to $50 million (USD)"         ~ "25000000-50000000",
            Revenue=="$50 to $100 million (USD)"        ~ "50000000-100000000",
            Revenue=="$100 to $500 million (USD)"       ~ "100000000-500000000",
            Revenue=="$500 million to $1 billion (USD)" ~ "500000000-1000000000",
            Revenue=="$1 to $2 billion (USD)"           ~ "1000000000-2000000000",
            Revenue=="$2 to $5 billion (USD)"           ~ "2000000000-5000000000",
            Revenue=="$5 to $10 billion (USD)"          ~ "5000000000-10000000000",
            Revenue=="$10+ billion (USD)"               ~ "10000000000-Inf",
        )

) %>%
  separate(Revenue, sep="-", into=c("Revenue Inf","Revenue Sup")) %>%
  mutate(
    `Revenue Inf`=as.numeric(`Revenue Inf`),
    `Revenue Sup`=as.numeric(`Revenue Sup`)
  )
```

Tratamiento de *outliers* de la variable `Salary Estimate Med`

```
boxplot(clean_data$`Salary Estimate Med`, main="Salario estimado promedio", col="blue")
```



```
salary_outliers <- unique(boxplot.stats(clean_data$`Salary Estimate Med`)$out)
salary_outliers
```

```
## [1]  43.5 271.5
```

```
clean_data %>%
  filter(
    `Salary Estimate Med` == 43.5
  )
```

```
## # A tibble: 17 x 22
##    index `Job Title`  `Salary Estimat~ `Salary Estimat~ `Job Description` Rating
##    <dbl> <chr>                   <dbl>            <dbl> <chr>              <dbl>
##  1   467 machine lea~               31               56 "Passionate abou~    3.3
##  2   468 data scient~               31               56 "As a member of ~    4
##  3   471 data scient~               31               56 "The Perduco Gro~    3.9
##  4   472 data engine~               31               56 "Western Digital~    3.5
##  5   473 data scient~               31               56 "Data Scientist\~    4
##  6   474 data scient~               31               56 "Deepen understa~    3
##  7   475 data scient~               31               56 "Job Introductio~    3.7
##  8   476 data scient~               31               56 "Job Title : Dat~    4.5
##  9   477 data scient~               31               56 "Child Care Awar~    2.8
## 10   478 data scient~               31               56 "Responsibilitie~    4.1
## 11   479 data analyst               31               56 "In-Line Inspect~    2.9
## 12   480 data scient~               31               56 "What will you b~    4.7
## 13   481 data scient~               31               56 "General Descrip~    3.4
## 14   482 data scient~               31               56 "Job Description~    3.9
## 15   483 data scient~               31               56 "Company overvie~    2.9
## 16   485 data scient~               31               56 "Join us in maki~    4.9
## 17   486 data scient~               31               56 "US Citizenship ~    5
## # ... with 16 more variables: Company Name <chr>, Location <chr>,
## #   Headquarters <chr>, Size Inf <dbl>, Size Sup <dbl>, Founded <dbl>,
## #   Type of ownership <chr>, Industry <chr>, Sector <chr>, Revenue Inf <dbl>,
## #   Revenue Sup <dbl>, Competitors <chr>, Salary Estimate Med <dbl>,
## #   Same Location Headquarter <lgl>, Size Ordered <dbl>, Revenue Ordered <dbl>
```

```
clean_data %>%
  filter(
    `Salary Estimate Med` == 271.5
  )
```

```
## # A tibble: 16 x 22
##    index `Job Title`  `Salary Estimat~ `Salary Estimat~ `Job Description` Rating
##    <dbl> <chr>                   <dbl>            <dbl> <chr>              <dbl>
##  1   508 data scient~              212              331 "Roche Diagnosti~    4.1
##  2   509 data scient~              212              331 "Title: Real Wor~    4
##  3   510 data scient~              212              331 "Position: Data ~    3.6
##  4   511 data scient~              212              331 "Company: AI/Dat~    5
##  5   512 data scient~              212              331 "Do you have a h~    3.5
##  6   513 data scient~              212              331 "NO OPT CPT pls.~    4.7
##  7   514 data scient~              212              331 "Please review t~    3.5
##  8   515 data scient~              212              331 "Ke`aki Technolo~    3.6
##  9   517 data scient~              212              331 "JOB DESCRIPTION~    2.7
## 10   520 data scient~              212              331 "Role: Data Scie~    4
## 11   521 data scient~              212              331 "The work is in ~    3.8
```

```
## 12   524 data scient~              212              331 "CompuForce is s~   NA
## 13   525 data scient~              212              331 "Job Description~    4.5
## 14   526 machine lea~              212              331 "Scientist 1 - M~    3.5
## 15   527 data scient~              212              331 "Description\r\n~    2.7
## 16   528 data scient~              212              331 "Aptive is seeki~    3.5
## # ... with 16 more variables: Company Name <chr>, Location <chr>,
## #   Headquarters <chr>, Size Inf <dbl>, Size Sup <dbl>, Founded <dbl>,
## #   Type of ownership <chr>, Industry <chr>, Sector <chr>, Revenue Inf <dbl>,
## #   Revenue Sup <dbl>, Competitors <chr>, Salary Estimate Med <dbl>,
## #   Same Location Headquarter <lgl>, Size Ordered <dbl>, Revenue Ordered <dbl>
```

Análisis

Estudio de la normalidad de variable `Salary Estimate Med`

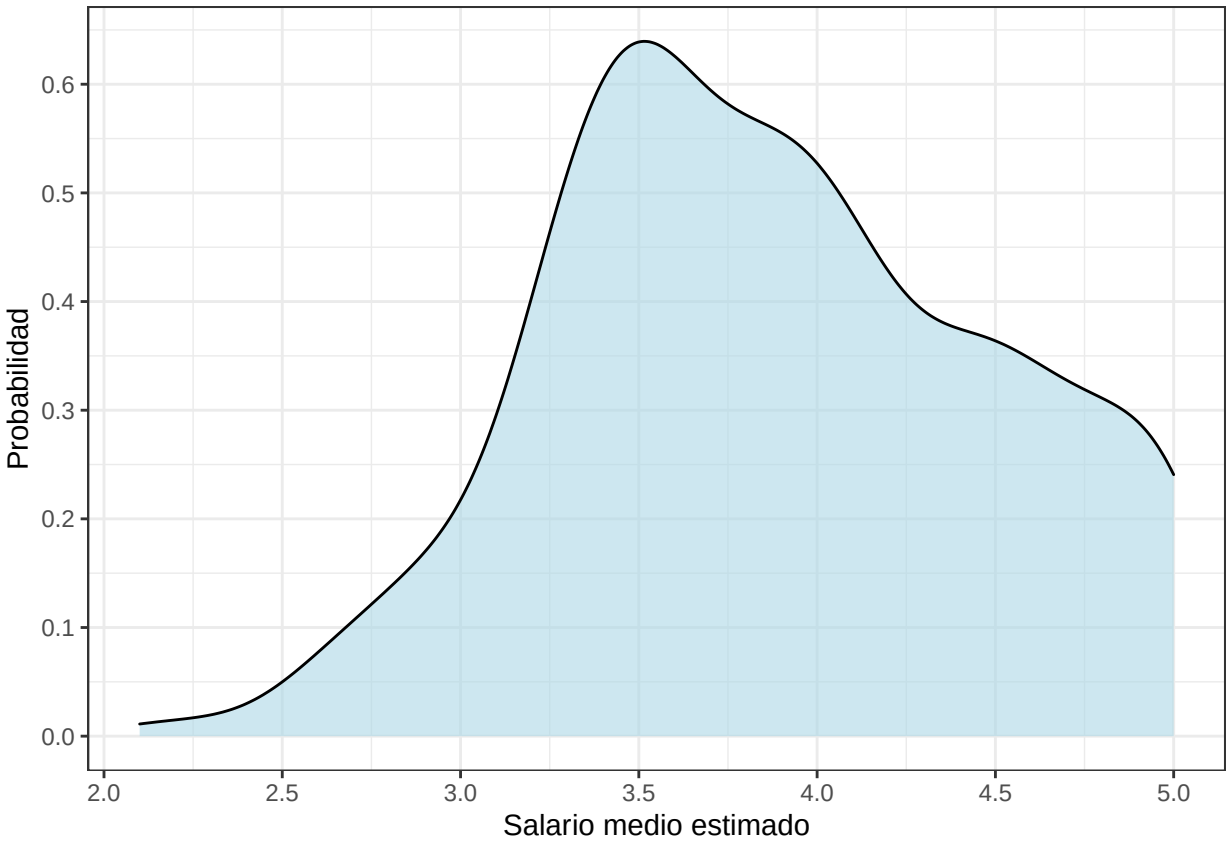
```
n_bins=clean_data %>%
  distinct(`Salary Estimate Med`) %>%
  as_vector() %>%
  length()
```

```
histogram_salary <- clean_data %>%
  ggplot(aes(x=`Salary Estimate Med`)) +
  geom_histogram(bins=n_bins, fill="lightblue",color="black") +
  theme_bw() +
  scale_x_continuous(breaks=scales::pretty_breaks(n = 10)) +
  scale_y_continuous(breaks=scales::pretty_breaks(n = 10)) +
  labs(x="Salario medio estimado", y="Conteo")
```

```
density_salary <- clean_data %>%
  ggplot(aes(x=`Salary Estimate Med`)) +
  geom_density(alpha=0.6, fill="lightblue") +
  theme_bw() +
  scale_x_continuous(breaks=scales::pretty_breaks(n = 10)) +
  scale_y_continuous(breaks=scales::pretty_breaks(n = 10)) +
  labs(x="Salario medio estimado", y="Probabilidad")
```

```
normality_test_salary <- shapiro.test(clean_data$`Salary Estimate Med`)
```

```
clean_data %>%
  ggplot(aes(x=`Rating`)) +
  geom_density(alpha=0.6, fill="lightblue") +
  theme_bw() +
  scale_x_continuous(breaks=scales::pretty_breaks(n = 10)) +
  scale_y_continuous(breaks=scales::pretty_breaks(n = 10)) +
  labs(x="Salario medio estimado", y="Probabilidad")
```



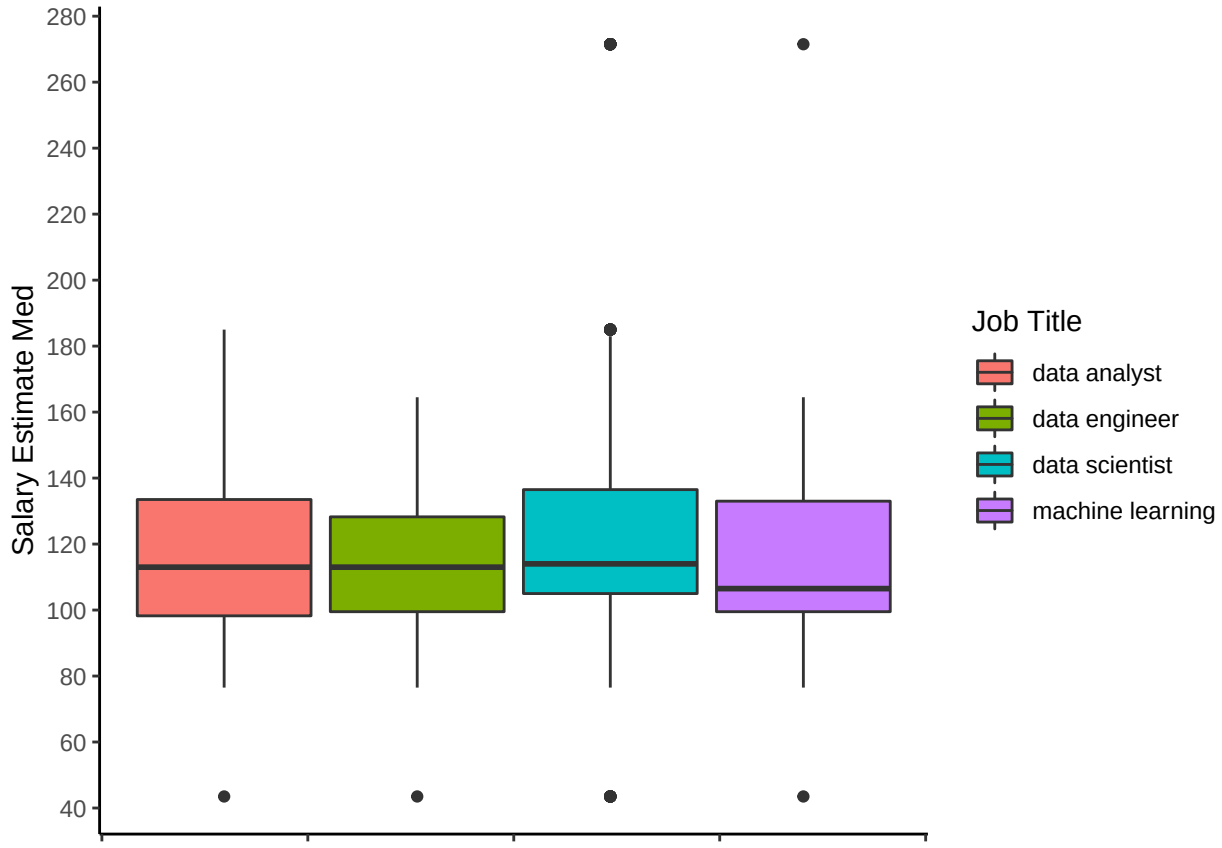

```
shapiro.test(clean_data$Rating)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  clean_data$Rating
## W = 0.97766, p-value = 2.529e-07
```

Estudio de relación entre `Estimate Salary Med` y otras variables

Categóricas: `Job Title`, `Size Ordered`, `Revenue Ordered` y `Industry`

```
clean_data %>%
  ggplot(aes(y=`Salary Estimate Med`, fill=`Job Title`)) +
  scale_x_continuous(labels=NULL) +
  scale_y_continuous(breaks=scales::pretty_breaks(n = 10)) +
  theme_classic() +
  geom_boxplot()
```




```
kruskal.test(`Salary Estimate Med` ~ `Job Title`, data = clean_data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Salary Estimate Med by Job Title
## Kruskal-Wallis chi-squared = 6.0209, df = 3, p-value = 0.1106
```

```
kruskal.test(`Salary Estimate Med` ~ `Size Ordered`, data = clean_data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Salary Estimate Med by Size Ordered
## Kruskal-Wallis chi-squared = 6.0905, df = 6, p-value = 0.4131
```

```
kruskal.test(`Salary Estimate Med` ~ `Revenue Ordered`, data = clean_data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Salary Estimate Med by Revenue Ordered
## Kruskal-Wallis chi-squared = 17.603, df = 11, p-value = 0.09127
```

```
# Estudio de las diferentes industrias en el dataset y sus concurrencias
sort(table(clean_data$Industry),descending=T)
```

```
##
##    Cable, Internet & Telephone Providers
```

```
##                                     1
##      Department, Clothing, & Shoe Stores
##                                     1
##                    Farm Support Services
##                                     1
##                   Food & Beverage Stores
##                                     1
##                Hotels, Motels, & Resorts
##                                     1
##                 Logistics & Supply Chain
##                                     1
##                               News Outlet
##                                     1
##                                Publishing
##                                     1
##                                     Rail
##                                     1
##                          Social Assistance
##                                     1
##                State & Regional Agencies
##                                     1
##         Telecommunications Manufacturing
##                                     1
##                Transportation Management
##                                     1
##                   Chemical Manufacturing
##                                     2
##                  Colleges & Universities
##                                     2
##                              Construction
##                                     2
## Consumer Electronics & Appliances Stores
##                                     2
##                Express Delivery Services
##                                     2
##         Financial Transaction Processing
##                                     2
##                Health, Beauty, & Fitness
##                                     2
##                        Oil & Gas Services
##                                     2
##                         Timber Operations
##                                     2
##   Transportation Equipment Manufacturing
##                                     2
##                           Travel Agencies
##                                     2
##                                Utilities
##                                     2
##         Venture Capital & Private Equity
##                                     2
##                                Accounting
##                                     3
##     Architectural & Engineering Services
```

```
##                                     3
##          Consumer Products Manufacturing
##                                     3
##    Electrical & Electronic Manufacturing
##                                     3
##                Industrial Manufacturing
##                                     3
##                     Other Retail Stores
##                                     3
##                              Real Estate
##                                     3
##                              Video Games
##                                     3
##                                Wholesale
##                                     3
##            Food & Beverage Manufacturing
##                                     4
##          Insurance Agencies & Brokerages
##                                     4
##                                  Energy
##                                     5
##              Telecommunications Services
##                                     5
##                                 Lending
##                                     7
##                    Banks & Credit Unions
##                                     8
##                   Research & Development
##                                     8
##    Investment Banking & Asset Management
##                                    11
##                          Federal Agencies
##                                    15
##         Health Care Services & Hospitals
##                                    15
##                  Advertising & Marketing
##                                    22
##                       Insurance Carriers
##                                    22
##                                Internet
##                                    26
##                              Consulting
##                                    36
##                   Staffing & Outsourcing
##                                    36
##                      Aerospace & Defense
##                                    38
##  Enterprise Software & Network Solutions
##                                    39
##                Biotech & Pharmaceuticals
##                                    43
##             Computer Hardware & Software
##                                    50
##                              IT Services
```

```
##                                     58
```

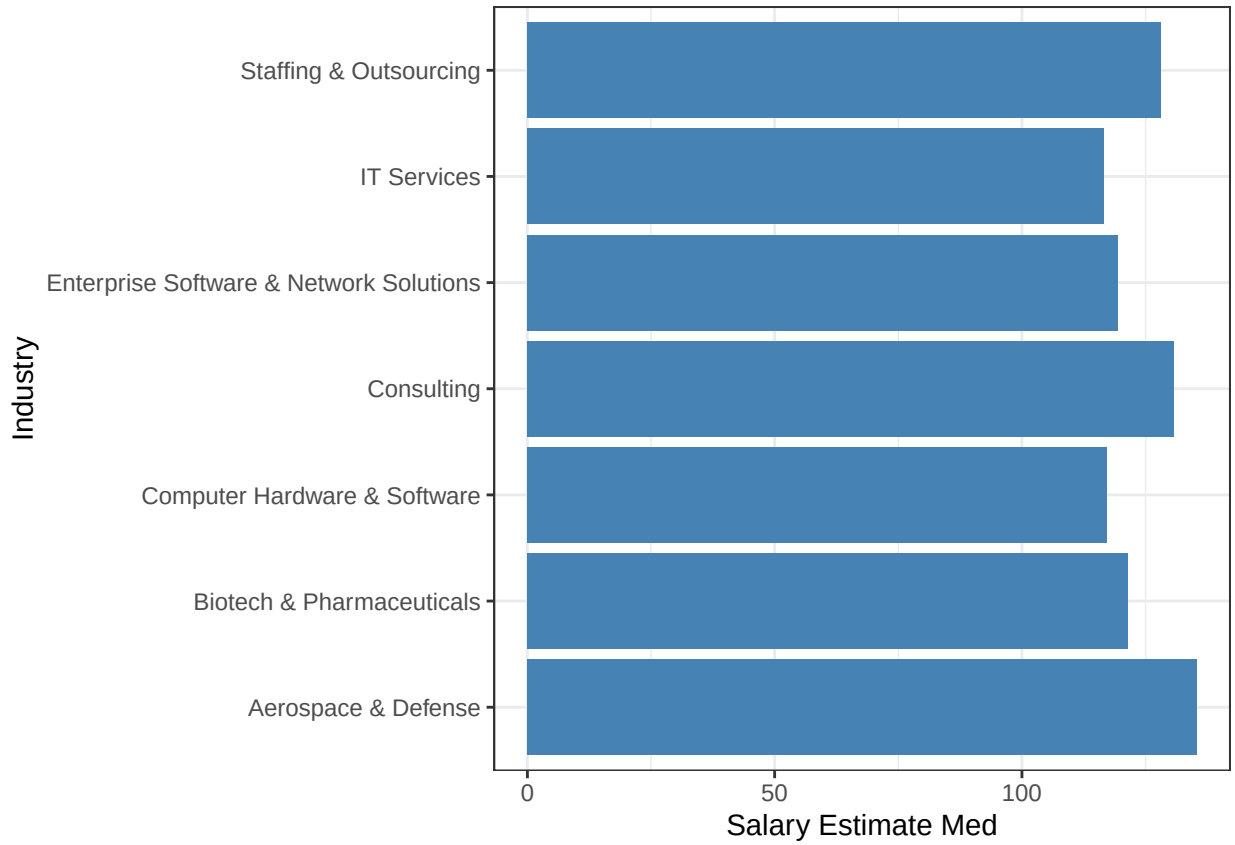
```
length(unique(clean_data$Industry))
```

```
## [1] 56
```

```
clean_data %>%
  filter(!is.na(Industry)) %>%
  count(Industry, sort=T)
```

```
## # A tibble: 55 x 2
##    Industry                                  n
##    <chr>                                 <int>
##  1 IT Services                              58
##  2 Computer Hardware & Software             50
##  3 Biotech & Pharmaceuticals                43
##  4 Enterprise Software & Network Solutions  39
##  5 Aerospace & Defense                      38
##  6 Consulting                               36
##  7 Staffing & Outsourcing                   36
##  8 Internet                                 26
##  9 Advertising & Marketing                  22
## 10 Insurance Carriers                       22
## # ... with 45 more rows
```

```
clean_data %>%
  filter(!is.na(Industry)) %>%
  add_count(Industry) %>%
  arrange(desc(n)) %>% # igual a add_count(Industry,sort=T)
  slice_max(n, prop=0.5) %>%
  ggplot(aes(x=Industry, y=`Salary Estimate Med`)) + geom_bar(stat="summary", fun="mean", fill="steelblu
```

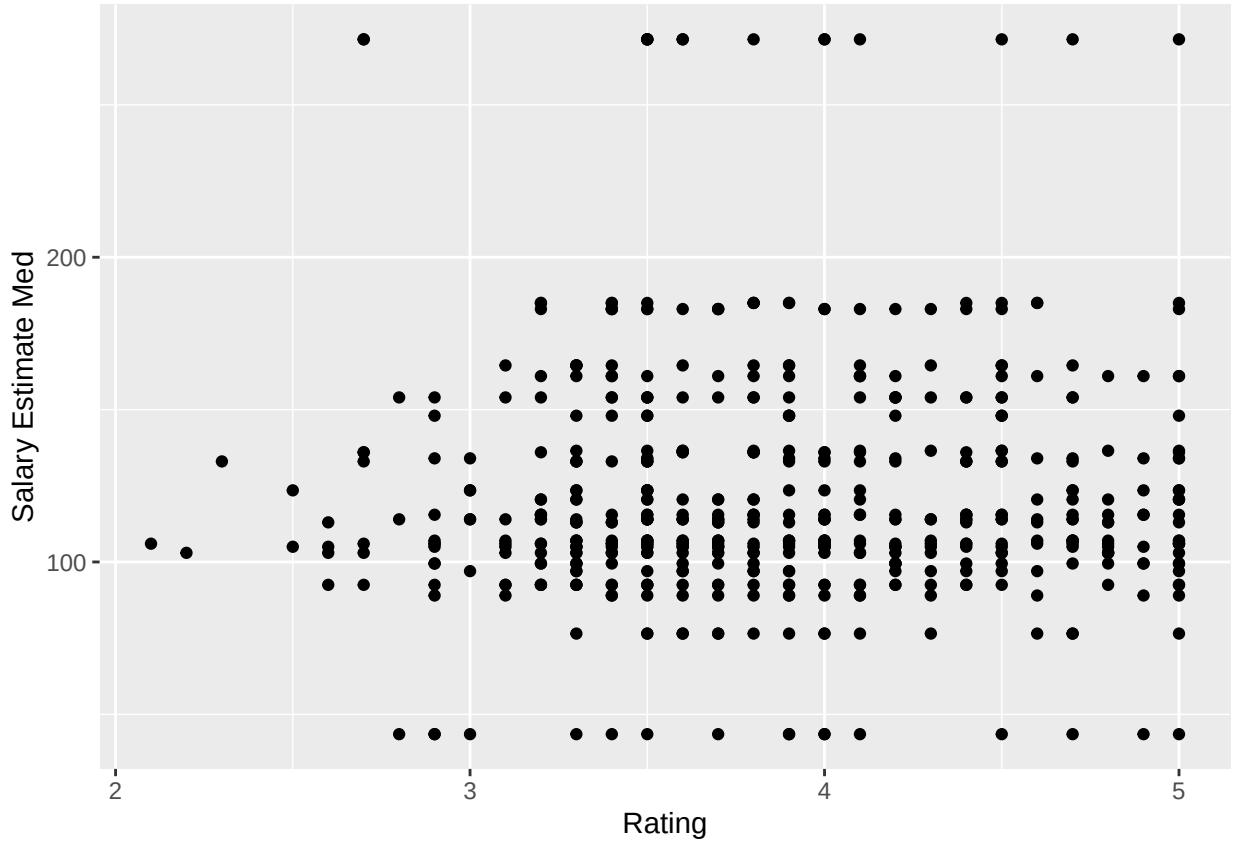



```
kruskal.test(`Salary Estimate Med` ~ Industry, data = clean_data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Salary Estimate Med by Industry
## Kruskal-Wallis chi-squared = 65.675, df = 54, p-value = 0.1325
```

Continuas: Rating

```
clean_data %>%
  ggplot(aes(x=Rating,y=`Salary Estimate Med`)) +
  geom_point()
```




```
cor(clean_data$Rating, clean_data$`Salary Estimate Med`, use="complete.obs")
```

```
## [1] 0.004244959
```

Binarias: Same Location Headquarter

```
summary(glm(`Same Location Headquarter`~`Salary Estimate Med`,family=binomial(link=logit), data=clean_da
```

```
##
## Call:
## glm(formula = `Same Location Headquarter` ~ `Salary Estimate Med`,
##     family = binomial(link = logit), data = clean_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0678  -0.9959  -0.9445   1.3679   1.5794
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -0.140313   0.292857  -0.479    0.632
## `Salary Estimate Med` -0.002830   0.002293  -1.234    0.217
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 740.10  on 556  degrees of freedom
## Residual deviance: 738.54  on 555  degrees of freedom
```

```
##   (28 observations deleted due to missingness)
## AIC: 742.54
##
## Number of Fisher Scoring iterations: 4
```