Práctica 2

Rubén Silva Marín

30/5/2020

Contents

1. DESCRIPCIÓN DEL DATASET

```
library('ggplot2')
library('car')
```

```
## Warning: package 'car' was built under R version 3.6.3
```

```
## Loading required package: carData
```

```
setwd('G:\\Mi unidad\\UOC\\Data Science\\2\\Tipología y Ciclo de Vida de los Datos\\Práctica 2\\1632_774
data <- read.csv('Datos_Originales.csv')
```

El dataset contiene información (ficticia) de empleados de una empresa. Los datos son tanto demográficos como laborales. Así, es posible encontrar los siguientes campos:

```
summary(data)
```

```
##                   Employee_Name     EmpID             MarriedID
##                           : 91   Min.   :6.020e+08   Min.   :0.0000
##  Adinolfi, Wilson  K      :  1   1st Qu.:1.101e+09   1st Qu.:0.0000
##  Ait Sidi, Karthikeyan    :  1   Median :1.203e+09   Median :0.0000
##  Akinkuolie, Sarah        :  1   Mean   :1.200e+09   Mean   :0.3968
##  Alagbe,Trina             :  1   3rd Qu.:1.379e+09   3rd Qu.:1.0000
##  Albert, Michael          :  1   Max.   :1.988e+09   Max.   :1.0000
##  (Other)                  :305   NA's   :91          NA's   :91
##  MaritalStatusID     GenderID      EmpStatusID        DeptID
##  Min.   :0.0000   Min.   :0.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:1.000   1st Qu.:5.000
##  Median :1.0000   Median :0.000   Median :1.000   Median :5.000
##  Mean   :0.8097   Mean   :0.429   Mean   :2.397   Mean   :4.606
##  3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:5.000   3rd Qu.:5.000
##  Max.   :4.0000   Max.   :1.000   Max.   :5.000   Max.   :6.000
##  NA's   :91       NA's   :91      NA's   :91      NA's   :91
##   PerfScoreID    FromDiversityJobFairID    PayRate          Termd
##  Min.   :1.000   Min.   :0.00000        Min.   :14.00   Min.   :0.0000
```

```
##  1st Qu.:3.000   1st Qu.:0.00000        1st Qu.:20.00   1st Qu.:0.0000
##  Median :3.000   Median :0.00000        Median :24.00   Median :0.0000
##  Mean   :2.984   Mean   :0.09355        Mean   :31.28   Mean   :0.3323
##  3rd Qu.:3.000   3rd Qu.:0.00000        3rd Qu.:45.31   3rd Qu.:1.0000
##  Max.   :4.000   Max.   :1.00000        Max.   :80.00   Max.   :1.0000
##  NA's   :91      NA's   :91             NA's   :91      NA's   :91
##    PositionID                         Position       State          Zip
##  Min.   : 1.00   Production Technician I :136   MA     :275   Min.   : 1013
##  1st Qu.:18.00                           : 91          : 91   1st Qu.: 1901
##  Median :19.00   Production Technician II: 57   CT     :  6   Median : 2132
##  Mean   :16.84   Area Sales Manager      : 27   TX     :  3   Mean   : 6570
##  3rd Qu.:20.00   Production Manager      : 14   VT     :  2   3rd Qu.: 2357
##  Max.   :30.00   Software Engineer       :  9   AL     :  1   Max.   :98052
##  NA's   :91      (Other)                 : 67   (Other): 23   NA's   :91
##        DOB      Sex         MaritalDesc               CitizenDesc
##          : 91     : 91            : 91                      : 91
##  06/14/87:  2   F :177   Divorced : 30   Eligible NonCitizen: 12
##  07/07/84:  2   M :133   Married  :123   Non-Citizen        :  4
##  09/09/65:  2            Separated: 12   US Citizen         :294
##  09/22/76:  2            Single   :137
##  01/02/51:  1            Widowed  :  8
##  (Other) :301
##  HispanicLatino                                RaceDesc       DateofHire
##     : 91                                          : 91            : 91
##  no :  1         American Indian or Alaska Native:  4   1/10/2011: 14
##  No :281         Asian                            : 34   3/30/2015: 12
##  yes:  1         Black or African American        : 57   1/5/2015 : 11
##  Yes: 27         Hispanic                         :  4   9/29/2014: 11
##                  Two or more races                : 18   5/16/2011: 10
##                  White                            :193   (Other)  :252
##  DateofTermination                    TermReason
##          :298      N/A - still employed     :196
##  01/09/12:  2                               : 92
##  04/01/13:  2      Another position         : 20
##  04/04/14:  2      unhappy                  : 14
##  04/07/12:  2      more money               : 11
##  05/01/16:  2      N/A - Has not started yet: 11
##  (Other) : 93      (Other)                  : 57
##                EmploymentStatus                Department          ManagerName
##                       : 91                          : 91                 : 91
##  Active               :182     Admin Offices       : 10   Elijiah Gray  : 22
##  Future Start         : 11     Executive Office    :  1   Kelley Spirea : 22
##  Leave of Absence     : 14     IT/IS               : 50   Kissy Sullivan: 22
##  Terminated for Cause : 15     Production          :208   Michael Albert: 22
##  Voluntarily Terminated: 88    Sales               : 31   Amy Dunn      : 21
##                                Software Engineering: 10   (Other)       :201
##    ManagerID                              RecruitmentSource
##  Min.   : 1.00                                    : 91
##  1st Qu.:10.00   Employee Referral                 : 31
##  Median :15.00   Diversity Job Fair                : 29
##  Mean   :14.58   Search Engine - Google Bing Yahoo: 25
##  3rd Qu.:19.00   Monster.com                       : 24
##  Max.   :39.00   Pay Per Click - Google            : 21
##  NA's   :99      (Other)                           :180
```

```
##           PerformanceScore EngagementSurvey EmpSatisfaction
##                  : 91      Min.   :1.030    Min.   :1.00
##  Exceeds         : 37      1st Qu.:2.083    1st Qu.:3.00
##  Fully Meets     :243      Median :3.470    Median :4.00
##  Needs Improvement: 18     Mean   :3.332    Mean   :3.89
##  PIP             : 12      3rd Qu.:4.520    3rd Qu.:5.00
##                            Max.   :5.000    Max.   :5.00
##                            NA's   :91       NA's   :91
##  SpecialProjectsCount LastPerformanceReview_Date DaysLateLast30
##  Min.   :0.00                  :194              Min.   :0
##  1st Qu.:0.00         1/14/2019: 18              1st Qu.:0
##  Median :0.00         2/18/2019: 12              Median :0
##  Mean   :1.21         1/21/2019: 10              Mean   :0
##  3rd Qu.:0.00         1/28/2019:  9              3rd Qu.:0
##  Max.   :8.00         2/25/2019:  9              Max.   :0
##  NA's   :91           (Other)  :149              NA's   :194
```

- Employee_name: Indica el nombre de la persona (es ficticio)
- EmpID: el identificador de la persona (único)
- MarriedID: si la persona está casada (1) o no (0)
- MaritalStatusID: indica el codigo del estado civil de la persona: soltero (0), casado (1), divorciado (2), separado (3) o viudo (4)
- GenderID: indica el sexo de la persona: femenino (0) o masculino (1)
- EmpStatusID: indica el código del estado de empleo de la persona: activo (1), proxima contratación (2), excedencia (3), despedido por alguna causa (4) o baja voluntaria (5)
- DeptID: número de identificación del dpto.: oficinas administrativas (1), oficina ejecutiva (2), departamento de informática (3), desarrollo de software (4), producción (5) y ventas (6)
- PerfScoreID: número de identificación de la calificación de desempeño: necesita un plan de mejora (1), necesita mejorar (2), cumple objetivos (3), excede objetivos (4)
- FromDiversityJobFairID: indica si el empleado fue reclutado a través de alguna feria de empleo de diversidad: si (1) o no (0)
- PayRate: el salario por hora de la persona
- Termd: indica si el empleado ha sido despedido o no: si (1) o no (0)
- PositionID: indica el código del puesto de trabajo de la persona: contable (1), asistente administrativo (2), coordinador de zona de venta (3), desarrollador Business Intelligence (4), director de Business Intelligence (5), director ejecutivo de información (6), arquitecto de datos (7), administrador de bases de datos (8), analista de datos (9), director de operaciones (10), director de ventas (11), director de informática (12), manager de IT (13), soporte de IT (14), ingeniero de redes (15), CEO (16), manager de producción (17), técnico de producción I (18), técnico de producción II (19), jefe de ventas (20), desarrollador de business intelligence senior (21), manager de servicios compartidos (22), ingeniero de software (23), manager de ingenieros de software (24), contable senior (25), administrador de bases de datos senior (26), ingeniero de redes senior (27), arquitecto de datos principal (28) y arquitecto empresarial (29)
- Position: indica el nombre de la posición
- State: indica el estado en el que vive la persona
- Zip: indica el código postal de la ciudad en la que vive la persona
- DOB: indica la fecha de nacimiento de la persona
- Sex: indica el sexo de la persona: F para mujer y M para hombre
- MaritalDesc: indica el estado civil de la persona: solterx, casadx, separadx, divorciadx o viudx
- CitizenDesc: indica si la persona es ciudadana de los estados unidos: ciudadano o ciudadano extranjero
- HispanicLatino: si la persona es (1) o no (0) latina
- RaceDesc: raza de la persona: negro o afroamericano, blanco, asiático, nativo americano, hispano o mestizo de varias razas
- DateofHire: fecha de contratación

- DateofTermination: fecha de salida de la compañía
- TermReason: motivo de la salida: aún trabajando, cambio de carrera, cambio de posición, ausencias, reubicación fuera del área, no ha empezado aún, desempeño, no se presenta, horas, motivos médicos, jubilación, insatisfacción, mejor salario, vuelve a formación, conducta inapropiada, servicio militar o no vuelve de la baja por maternidad
- EmploymentStatus: indica el estado laboral de la persona: activo, despedido por alguna causa, baja voluntaria, próximo inicio o excedencia
- Departament: indica el nombre del departamento al que pertenece (anteriormente mencionados)
- ManagerName: indica el nombre del manager del departamento al que pertenece
- ManagerID: indica el identificador del manager
- RecruitmentSource: indica la fuente de reclutamiento utilizado para seleccionar al empleado: feria de empleo de diversidad, anuncios en página web, búsqueda por internet, campañas de pago de google, red social de la empresa, valla publicitaria, a través de la página de monster.com, periódico/revista, círculos profesionales/networking, referencia de empleado interno, página web Indeed, capañas de SEO, página web Glassdoor, referencia de proveedor externo, anuncios de MBTA, sesión informativa, boca a boca, campañas de pago, reclutamiento in-situ (campus), página web de la empresa, página web Careerbuilder o partners de la compañía
- PerformanceScore: indica el desempeño del empleado (anteriormente mencionado)
- EngagementSurvey: indica la puntuación sobre el nivel de compromiso de un empleado con la empresa
- EmpSatisfaction: indica la satisfacción del empleado del 1 (poca) al 5 (mucha)
- SpecialProjectsCount: indica el número de proyectos especiales en los que el trabajador ha participado en los últimos 6 meses
- LastPerformanceReview_Date: la fecha más reciente en la que el empleado ha tenido una revisión del desempeño
- DaysLateLast30: el número de veces en las que el empleado ha llegado tarde en los últimos 30 días

1.2. Importancia y objetivos de los análisis Este dataset es muy interesante para poder realizar distintos análisis que vayan en busca de confirmar que la empresa tiene implantada o no una política de igualdad en cuanto a la participación en proyectos especiales o el salario. En la época en la que estamos es muy importante tener fuertes políticas para evitar sesgos a la hora de ofrecer igualdad de oportunidades de desarrollo y de compensación económica. Este análisis es posible hacerse a través del análisis de la varianza comparando distintas varibles como el salario o la igualdad de oportunidad en proyectos importantes en función del sexo y la raza.

También sería posible evaluar la relación que existe entre distintas posiciones y/o departamentos con el salario.

Una vez descartadas los sesgos negativos hacia la diversidad (con suerte) y revisadas las relaciones entre las posiciones y departamentos, se podría construir un modelo de regresión lineal del salario por hora en función de la participación en proyectos importantes la posición y el departamento.

2. INTEGRACIÓN Y SELECCIÓN DE LOS DATOS DE INTERÉS A ANALIZAR

Para llevar a cabo los análisis se tendrá que mantener aquellos datos que formarán parte del mismo. En este caso, se utilizarán las siguientes variables:
- Sexo | Sex
- Raza | HispanicLatino, RaceDesc
- Participación en proyectos importantes | SpecialProjectsCount
- Salario | PayRate
- Departamento | Department
- Posición | Position

Para poder seleccionarlas se volcarán en otra variable, dejando el origen intacto por si se necesita más adelante.

```
origen <- data
data <- data[, c('Sex', 'SpecialProjectsCount', 'PayRate', 'Department', 'Position', 'HispanicLatino',
```

Ahora se puede trabajar con el conjunto de variables seleccionado.

3. LIMPIEZA DE DATOS

3.1. ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos? Se van a analizar los datos para descubrir si tienen valores perdidos. Es posible comenzar con los valores 'NA' para ver si hay pérdidas en los datos.

```
colSums(is.na(data))
```

```
##                  Sex SpecialProjectsCount              PayRate
##                    0                   91                   91
##           Department             Position       HispanicLatino
##                    0                    0                    0
##             RaceDesc
##                    0
```

Como se puede observar, hay una pérdida sistemática de 91 datos en algunas de las variables. Por otro lado, es posible que también hayan 91 datos perdidos en estas variables que a priori no parece tenerlas.

```
head(data[is.na(data$PayRate),], 10)
```

```
##     Sex SpecialProjectsCount PayRate Department Position HispanicLatino
## 311                       NA      NA
## 312                       NA      NA
## 313                       NA      NA
## 314                       NA      NA
## 315                       NA      NA
## 316                       NA      NA
## 317                       NA      NA
## 318                       NA      NA
## 319                       NA      NA
## 320                       NA      NA
##     RaceDesc
## 311
## 312
## 313
## 314
## 315
## 316
## 317
## 318
## 319
## 320
```

Es posible ver que los valores perdidos en una de las variables que los presentaba van acompañados de valores perdidos en las otras variables. Además, se observa también que las variables que no presentaban a priori estos valores perdidos tienen valores vacíos, los cuales no tienen por qué corresponder con valores 'NA', dando negativo en el test de valores 'NA'.

```
colSums(data == '', na.rm = TRUE)
```

```
##                  Sex SpecialProjectsCount              PayRate
##                   91                    0                    0
##           Department             Position       HispanicLatino
##                   91                   91                   91
##             RaceDesc
##                   91
```

Se observa claramente que los 91 valores que anteriormente se observaban como perdidos se corresponden con valores vacíos en las variables que no presentaban esos 91 valores perdidos.

Se va a proceder a comprobar por último si existen otros valores posibles valores perdidos representados de alguna otra forma.

```
emptydata <- data.frame(colSums(data == '?', na.rm = TRUE), colSums(data == '999', na.rm = TRUE), colSu
colnames(emptydata) <- c('?', '999', '\' \'', '-', '0')
emptydata
```

```
##                       ? 999 ' ' -   0
## Sex                   0   0   0 0   0
## SpecialProjectsCount  0   0   0 0 242
## PayRate               0   0   0 0   0
## Department            0   0   0 0   0
## Position              0   0   0 0   0
## HispanicLatino        0   0   0 0   0
## RaceDesc              0   0   0 0   0
```

En la última columna, se comparan las variables con el posible valor perdido 0. Solo muestran 0 la variable SpecialProjectsCount, pero este es un valor del dominio posible y válido de esta. No sería un error o valor perdido.

En este caso, se ve claramente que los registros con valores perdidos deben ser eliminados porque no aportan ninguna información valiosa. Si se mantuviesen podrían suponer un problema y habría que imputar valores en todas las variables posibles, lo cual conlleva mucho esfuerzo para nada (los registros resultantes serían estimados en todas las variables).

```
data <- data[is.na(data$PayRate) == FALSE,]
colSums(is.na(data))
```

```
##                  Sex SpecialProjectsCount              PayRate
##                    0                    0                    0
##           Department             Position       HispanicLatino
##                    0                    0                    0
##             RaceDesc
##                    0
```

Ahora se observa que no hay valores perdidos en los datos.

Antes de continuar hacia el estudio de los valores extremos se va a comprobar si existen problemas en los datos como errores de formato o similares

```
str(data)
```

```
## 'data.frame':    310 obs. of  7 variables:
##  $ Sex                 : Factor w/ 3 levels "","F","M ": 2 3 3 2 2 2 2 2 2 3 ...
##  $ SpecialProjectsCount: int  6 4 5 4 5 4 0 0 0 0 ...
##  $ PayRate             : num  28.5 23 29 21.5 16.6 ...
##  $ Department          : Factor w/ 7 levels "","Admin Offices",..: 2 2 2 2 2 2 6 6 6 6 ...
##  $ Position            : Factor w/ 33 levels "","Accountant I",..: 2 2 2 3 3 3 4 4 4 4 ...
##  $ HispanicLatino      : Factor w/ 5 levels "","no","No","yes",..: 3 3 3 3 3 3 3 3 5 3 ...
##  $ RaceDesc            : Factor w/ 7 levels "","American Indian or Alaska Native",..: 4 4 7 7 7 3 4 7
```

Aquí se pueden observar varios aspectos:
- Es posible que haya errores en la variable sexo, ya que posee 3 niveles y el nivel M tiene un espacio después del carácter
- La variable HispanicLatino tiene errores claramente, ya que es una variable dicotómica y tiene 5 niveles
- Se comprobará el resto de variables cualitativas por si hay más errores

Sexo

```
levels(data$Sex)
```

```
## [1] ""   "F"  "M "
```

El tercer nivel observado anteriormente se eliminó con la limpieza de los valores perdidos. Posiblemente se haya quedado la configuración, aunque no haya valores erróneos. Se recodificará la variable para evitar errores.

```
data$Sex <- factor(data$Sex, levels = c('F', 'M '), labels = c('F', 'M'))
levels(data$Sex)
```

```
## [1] "F" "M"
```

Se han corregido los errores para que solo muestre los 2 valores posibles.

HispanicLatino

```
levels(data$HispanicLatino)
```

```
## [1] ""    "no"  "No"  "yes" "Yes"
```

Se observan 5 niveles que claramente no corresponden con los que debería tener. Uno vacío que se quedó aún después de eliminar los valores vacíos, y otros 4 valores que se repiten: 2 para sí y 2 para no.

```
data$HispanicLatino <- factor(data$HispanicLatino, levels = c('no', 'No', 'yes', 'Yes'), labels = c('No
levels(data$HispanicLatino)
```

```
## [1] "No"  "Yes"
```

Se han corregido los errores para que solo muestre los 2 valores posibles.

RaceDesk

```r
levels(data$RaceDesc)
```

```
## [1] ""                                "American Indian or Alaska Native"
## [3] "Asian"                           "Black or African American"
## [5] "Hispanic"                        "Two or more races"
## [7] "White"
```

Se observa un nivel vacío que se arrastra de cuando habían valores de este tipo. Se corregirá el factor para que solo contenga los niveles correctos.

```r
nivelesCorrectos <- levels(data$RaceDesc)[-1]
data$RaceDesc <- factor(data$RaceDesc, levels = nivelesCorrectos)
levels(data$RaceDesc)
```

```
## [1] "American Indian or Alaska Native" "Asian"
## [3] "Black or African American"        "Hispanic"
## [5] "Two or more races"                "White"
```

Ahora ya está corregido el factor.

Department

```r
levels(data$Department)
```

```
## [1] ""                     "Admin Offices"        "Executive Office"
## [4] "IT/IS"                "Production       "    "Sales"
## [7] "Software Engineering"
```

Tiene el mismo problema que las anteriores, arrastra el nivel vacío y además el nivel Producción tiene un error: hay muchos espacios después de la cadena.

```r
nivelesCorrectos <- levels(data$Department)[-1]
nivelesCorrectos[4] <- 'Production'
data$Department <- factor(data$Department, levels = levels(data$Department)[-1], labels = nivelesCorrec
levels(data$Department)
```

```
## [1] "Admin Offices"        "Executive Office"     "IT/IS"
## [4] "Production"           "Sales"                "Software Engineering"
```

Ya se ha corregido.

Position

```r
levels(data$Position)
```

```
##  [1] ""                           "Accountant I"
##  [3] "Administrative Assistant"   "Area Sales Manager"
##  [5] "BI Developer"               "BI Director"
##  [7] "CIO"                        "Data Analyst"
##  [9] "Data Analyst "              "Data Architect"
## [11] "Database Administrator"     "Director of Operations"
```

```
## [13] "Director of Sales"            "Enterprise Architect"
## [15] "IT Director"                  "IT Manager - DB"
## [17] "IT Manager - Infra"           "IT Manager - Support"
## [19] "IT Support"                   "Network Engineer"
## [21] "President & CEO"              "Principal Data Architect"
## [23] "Production Manager"           "Production Technician I"
## [25] "Production Technician II"     "Sales Manager"
## [27] "Senior BI Developer"          "Shared Services Manager"
## [29] "Software Engineer"            "Software Engineering Manager"
## [31] "Sr. Accountant"               "Sr. DBA"
## [33] "Sr. Network Engineer"
```

Esta variable, además de presentar el nivel vacío, tiene un nivel repetido y con un espacio extra: Data Analyst.

```
nivelesCorrectos <- levels(data$Position)
nivelesCorrectos[9] <- 'Data Analyst'
data$Position <- factor(data$Position, levels = levels(data$Position)[-1], labels = nivelesCorrectos[-1]
levels(data$Position)
```

```
##  [1] "Accountant I"                "Administrative Assistant"
##  [3] "Area Sales Manager"          "BI Developer"
##  [5] "BI Director"                 "CIO"
##  [7] "Data Analyst"                "Data Architect"
##  [9] "Database Administrator"      "Director of Operations"
## [11] "Director of Sales"           "Enterprise Architect"
## [13] "IT Director"                 "IT Manager - DB"
## [15] "IT Manager - Infra"          "IT Manager - Support"
## [17] "IT Support"                  "Network Engineer"
## [19] "President & CEO"             "Principal Data Architect"
## [21] "Production Manager"          "Production Technician I"
## [23] "Production Technician II"    "Sales Manager"
## [25] "Senior BI Developer"         "Shared Services Manager"
## [27] "Software Engineer"           "Software Engineering Manager"
## [29] "Sr. Accountant"              "Sr. DBA"
## [31] "Sr. Network Engineer"
```

Ahora se observa que los niveles están corregidos. Se eliminará la variable 'NivelesCorrectos' porque no tiene más utilidad.

```
rm(nivelesCorrectos)
```

Aquí se muestran las variables completas.

```
str(data)
```

```
## 'data.frame':    310 obs. of  7 variables:
##  $ Sex                 : Factor w/ 2 levels "F","M": 1 2 2 1 1 1 1 1 1 2 ...
##  $ SpecialProjectsCount: int  6 4 5 4 5 4 0 0 0 0 ...
##  $ PayRate             : num  28.5 23 29 21.5 16.6 ...
##  $ Department          : Factor w/ 6 levels "Admin Offices",..: 1 1 1 1 1 1 5 5 5 5 ...
##  $ Position            : Factor w/ 31 levels "Accountant I",..: 1 1 1 2 2 2 3 3 3 3 ...
##  $ HispanicLatino      : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 2 1 ...
##  $ RaceDesc            : Factor w/ 6 levels "American Indian or Alaska Native",..: 3 3 6 6 6 2 3 6 6
```

```r
summary(data[,sapply(data, is.numeric)])
```

3.2. Identificación y tratamiento de valores extremos

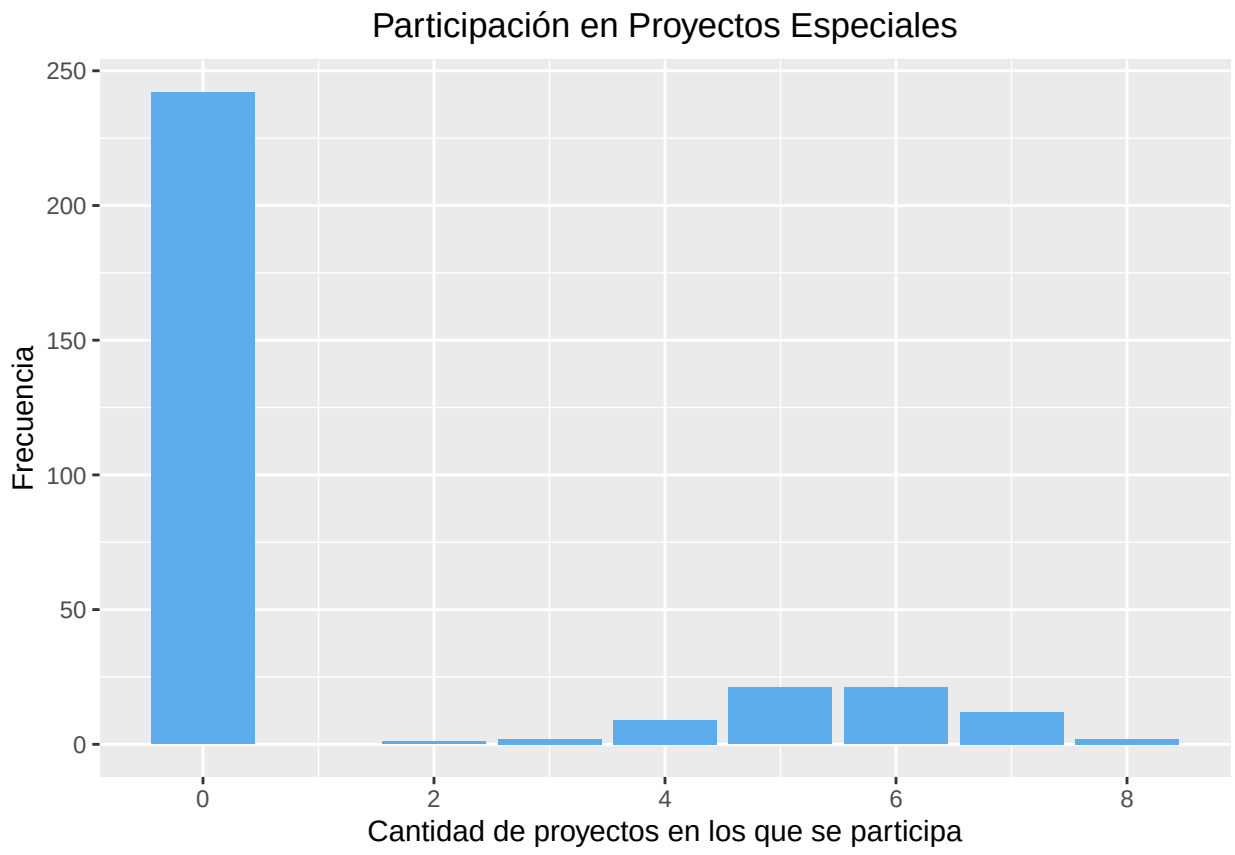
```
##  SpecialProjectsCount    PayRate
##  Min.   :0.00         Min.   :14.00
##  1st Qu.:0.00         1st Qu.:20.00
##  Median :0.00         Median :24.00
##  Mean   :1.21         Mean   :31.28
##  3rd Qu.:0.00         3rd Qu.:45.31
##  Max.   :8.00         Max.   :80.00
```

Es posible observar distintas características:
- la variable con mayor media es el salario, lo cual tiene sentido, dado que con respecto a las otras, su dominio es mayor
- la variable SpecialProjectsCount está centrada en 1, aunque su valor máximo es 8, lo cual se podría prever que está bastante desplazada hacia la izquierda (valores inferiores)
- la variable salario está centrada en 31, que es menos que la mitad del valor máximo, lo que podría esperarse que estuviera un poco desplazada hacia la izquierda

Se explorará la distribución de las variables numéricas:

SpecialProjectsCount

```r
ggplot(data, aes(SpecialProjectsCount)) + geom_bar(fill = 'steelblue2') + labs(title = 'Participación e
```



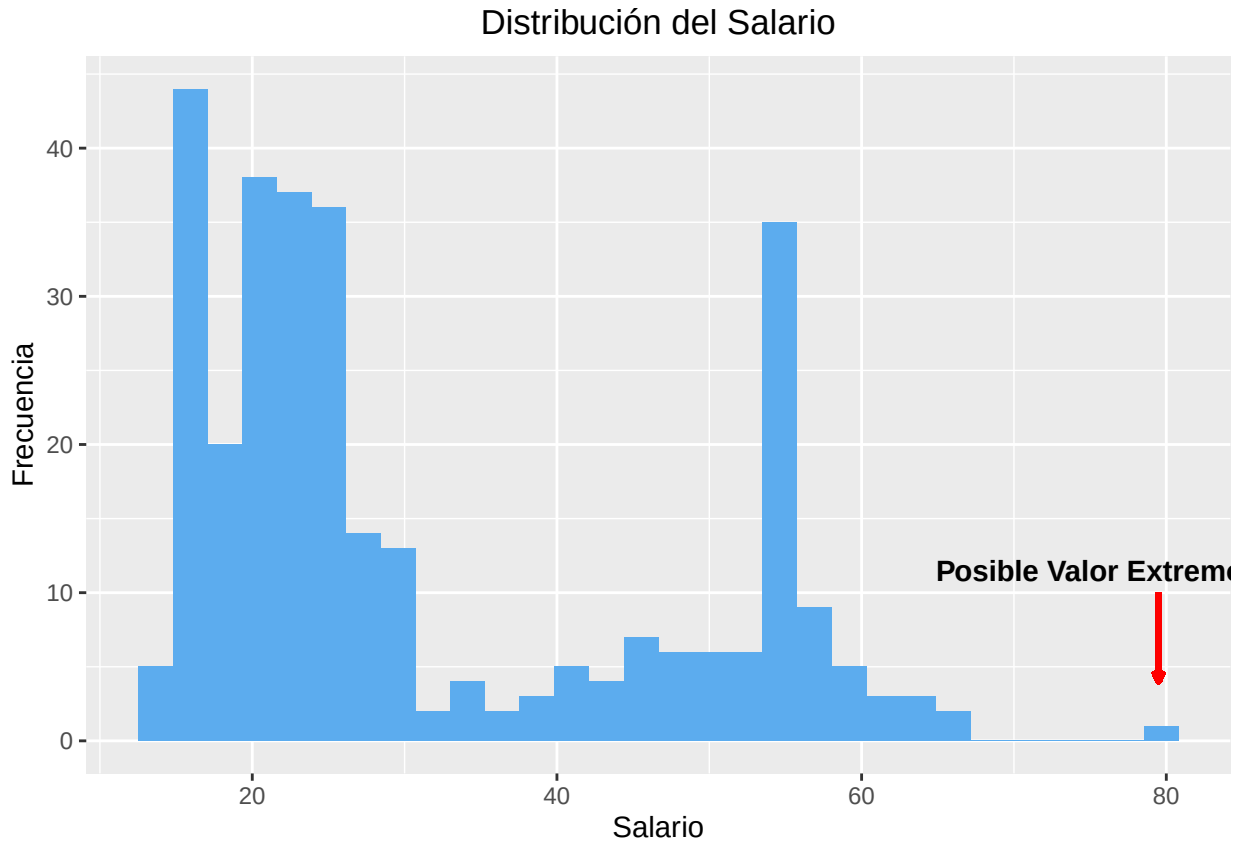
La distribución es asimétrica y parece tener sentido. La mayoría de los empleados no participan en un proyecto especial. Teniendo en cuenta que para ser especiales deberían ser limitados (no todos serán especiales) tiene sentido que no toda la plantilla participe. También se observa que es más probable que participe en varios proyectos que solo en uno.

Se observa que hay algunos empleados que participan hasta en 8 proyectos (valor máximo) pero no parece ser un valor extremo sin sentido, dado que hay varios empleados que participan en 8 proyectos y también hay muchos que participan en 7. No parece ser un error ni un valor que no se tenga que tener en cuenta.

PayRate

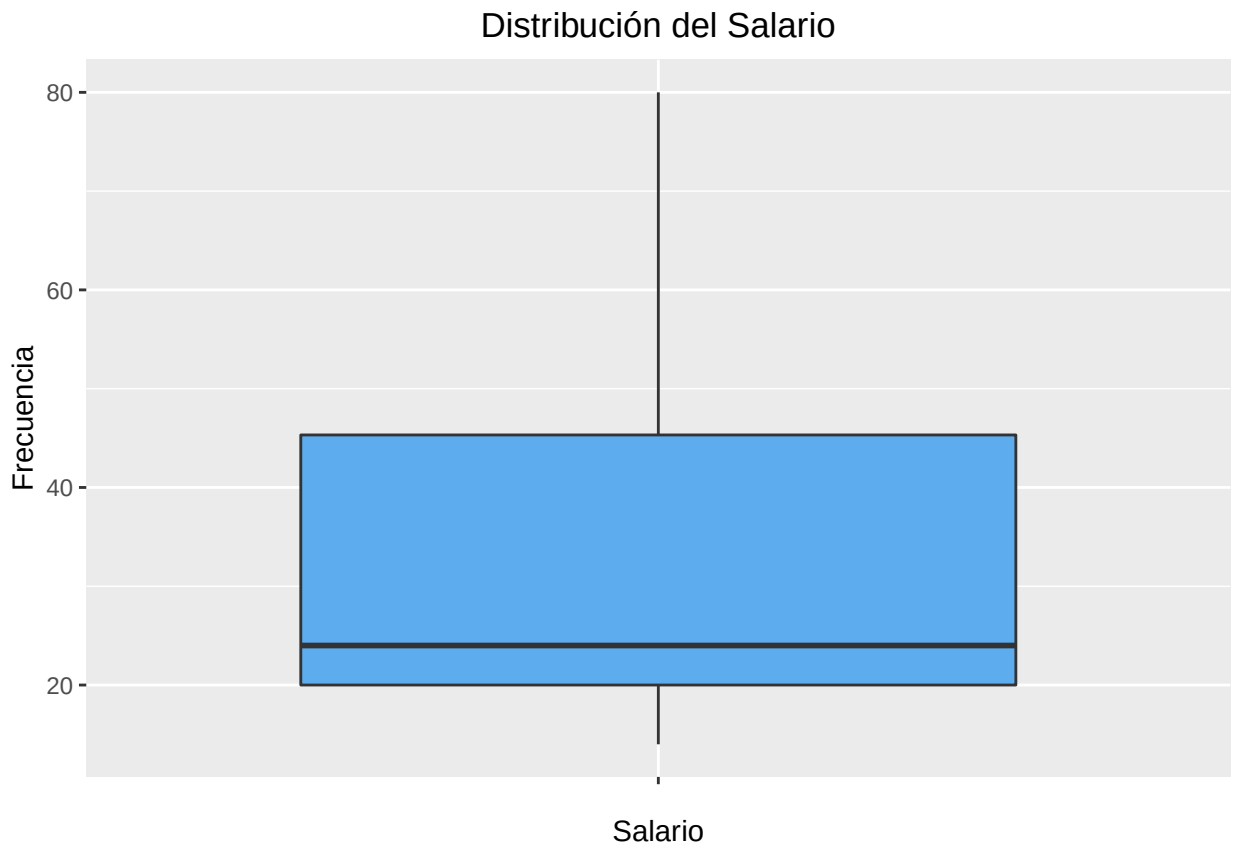
```
ggplot(data, aes(PayRate)) + geom_histogram(fill = 'steelblue2') + labs(title = 'Distribución del Salari
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



La distribución es asimétrica y no tiende a la normal. Hay muy pocas personas que cobren entre 30 y 55$/hora, mientras que hay muchos que cobran menos y también hay otro pico en torno a 55. Se observa un valor extremo situado en 80. Para poder estudiar si se debe mantener o no se comprobará si pertenece al CEO. En caso de que pertenezca, sería ideal eliminar esta figura del conjunto de datos porque es una figura especial que normalmente tiene otras condiciones laborales y situación distinta. Podría sesgar los análisis.

```
ggplot(data, aes(x = '', y = PayRate)) + geom_boxplot(fill = 'steelblue2') + labs(title = 'Distribución
```



No parece que en este caso el salario observado anteriormente sea un valor extremo. Sin embargo, tal y como se comentó anteriormente, las condiciones especiales del CEO se deberían considerar como distintas a las del resto de los empleados. Es por eso que se eliminará para no sesgar los análisis.

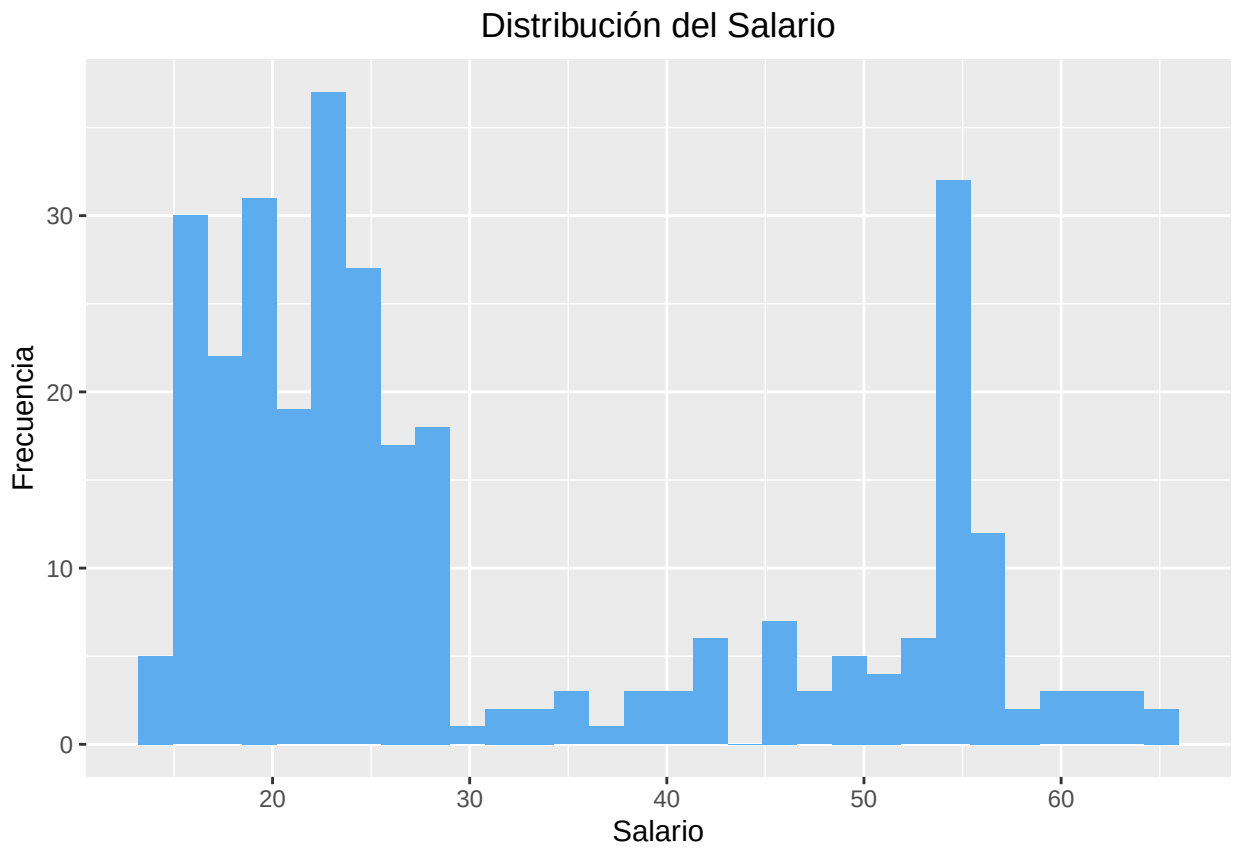
```
data[data$PayRate == 80,'Position']
```

```
## [1] President & CEO
## 31 Levels: Accountant I Administrative Assistant ... Sr. Network Engineer
```

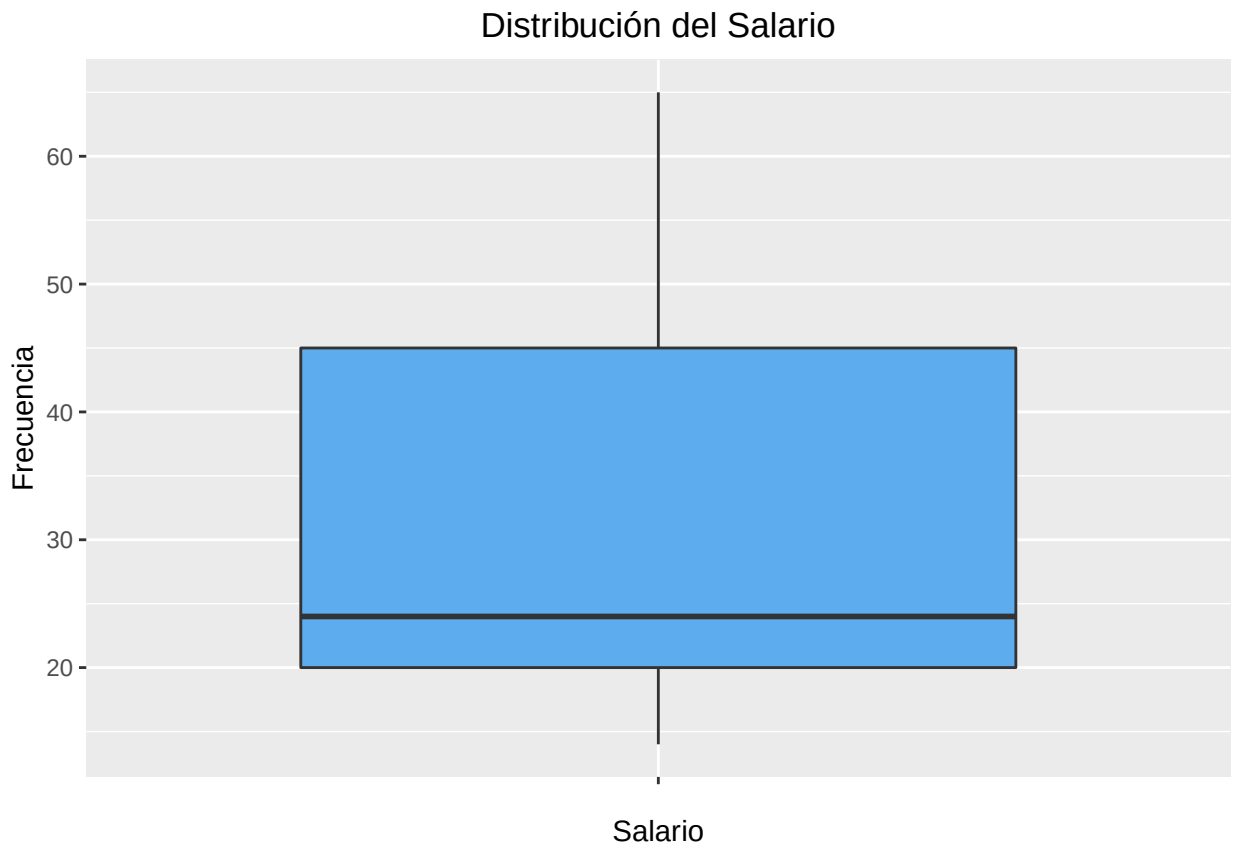
Se ve claramente que al ser la presidenta/CEO tiene mejores condiciones que el resto de empleados.

```
data <- data[data$PayRate != 80,]
ggplot(data, aes(PayRate)) + geom_histogram(fill = 'steelblue2') + labs(title = 'Distribución del Salari
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(data, aes(x = '', y = PayRate)) + geom_boxplot(fill = 'steelblue2') + labs(title = 'Distribución
```



Antes de continuar se comprobará que el departamento al que pertenecía la CEO no ha quedado vacío.

```
table(data$Department)
```

```
##
##        Admin Offices     Executive Office                IT/IS
##                   10                    0                   50
##           Production                Sales Software Engineering
##                  208                   31                   10
```

Como se puede observar se ha quedado sin ninguna persona. Se reasignarán los niveles del factor.

```
levels(data$Department)
```

```
## [1] "Admin Offices"        "Executive Office"     "IT/IS"
## [4] "Production"           "Sales"                "Software Engineering"
```

```
data$Department <- factor(data$Department, levels = c('Admin Offices', 'IT/IS', 'Production', 'Sales',
levels(data$Department)
```

```
## [1] "Admin Offices"        "IT/IS"                "Production"
## [4] "Sales"                "Software Engineering"
```

```r
table(data$Department)
```

```
##
##        Admin Offices                   IT/IS           Production
##                   10                      50                  208
##                Sales Software Engineering
##                   31                      10
```

Se observa que ya no hay valores extremos.

4. ANÁLISIS

4.1. Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar) Para llevar a cabo los análisis es necesario determinar qué tipo de pruebas se realizarán:
- Para comparar los salarios en función del sexo y la raza se podría realizar un análisis de la varianza. Por ejemplo, un ANOVA o un test no paramétrico de Mann Whitney o Kruskal-Wallis.
- Para comparar la participación en proyectos especiales en función del sexo o la raza se tendría que utilizar otro tipo de test dado que las variables dependientes es discreta. En este caso se utilizaría un test chi-cuadrado.
- Para el modelo de regresión se utilizará el método de los mínimos cuadrado y así intentar predecir el salario por hora en función de la participación en proyectos importantes, el departamento y/o la posición.

4.2. Comprobación de la normalidad y homogeneidad de la varianza. Como se han propuesto varios tipos de pruebas para el análisis del salario en función del sexo y la raza, es necesario analizar si la variable salario se ajusta a una distribución normal o no. Con esto se sabrá si se debe aplicar un test paramétrico o no paramétrico.

Salarios

```r
shapiro.test(data$PayRate)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$PayRate
## W = 0.8309, p-value < 2.2e-16
```

Se observa claramente que el salario no sigue una distribución normal. Lo que se sospechaba tras evaluar el histograma anterior.

Dado que no se puede aplicar el análisis de la varianza ANOVA porque el salario no sigue una distribución normal, habrá que aplicar el test de Mann Whitney para evaluar la relación con el sexo.

Por otro lado, para evaluar la relación del salario con la raza se debería aplicar el test no paramétrico de Kurskal-Wallis, dado que hay más de 2 categorías de raza. Sin embargo, hay poca representación de algunas razas por lo que se modificará la variable agrupando algunas de las razas para evaluarlas con un test no paramétrico de Mann Whitney.

4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos. Salario en función del Sexo

```r
summary(data$Sex)
```

```
##   F   M
## 176 133
```

Hipótesis:

H_0 : No existen diferencias significativas del salario en función del sexo
H_1 : Si existen diferencias significativas del salario en función del sexo

Nivel de confianza: 95%; $\alpha = 0.05$

```r
wilcox.test(data = data, PayRate ~ Sex)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  PayRate by Sex
## W = 10137, p-value = 0.04377
## alternative hypothesis: true location shift is not equal to 0
```

Como se observa, la probabilidad es menor a la establecida por el nivel de confianza ($\alpha = 0.05$) por lo que se rechaza la hipótesis nula y se considera que hay diferencias significativas en cuanto al salario en función del sexo. Una mala noticia dado que se están estudiando las políticas de igualdad.

Salario en función de la Raza

```r
table(data$RaceDesc)
```

```
##
## American Indian or Alaska Native                            Asian
##                                4                               34
##        Black or African American                         Hispanic
##                               57                                4
##                Two or more races                            White
##                               18                              192
```

Como se observa, hay muy pocos empleados de algunas razas, por lo que se va a comparar blancos vs resto para ver si hay diferencias. De esta forma, se tendrá que aplicar de nuevo el test de Mann Whitney en lugar del Kruskal-Wallis dado que no hay más de 2 factores en la variable independiente.

```r
data[data$RaceDesc == 'White', 'RaceDescDisc'] <- 'White'
data[data$RaceDesc != 'White', 'RaceDescDisc'] <- 'No-White'
data$RaceDescDisc <- factor(data$RaceDescDisc)
head(data[,c('RaceDesc', 'RaceDescDisc')])
```

```
##                    RaceDesc RaceDescDisc
## 1 Black or African American     No-White
## 2 Black or African American     No-White
## 3                     White        White
## 4                     White        White
## 5                     White        White
## 6                     Asian     No-White
```

```r
summary(data$RaceDescDisc)
```

```
## No-White    White
##      117      192
```

```r
str(data$RaceDescDisc)
```

```
##  Factor w/ 2 levels "No-White","White": 1 1 2 2 2 1 1 2 2 1 ...
```

Hipótesis:

H_0 : No existen diferencias significativas del salario en función de la raza
H_1 : Si existen diferencias significativas del salario en función de la raza

Nivel de confianza: 95%; $\alpha = 0.05$

```r
wilcox.test(data = data, PayRate ~ RaceDescDisc)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  PayRate by RaceDescDisc
## W = 12678, p-value = 0.05753
## alternative hypothesis: true location shift is not equal to 0
```

Como se observa, la probabilidad es mayor a la establecida por el nivel de confianza ($\alpha = 0.05$) por lo que no se rechaza la hipótesis nula y se considera que no hay diferencias significativas en cuanto al salario en función de la raza. Una buena noticia dado que se están estudiando las políticas de igualdad.

Participación en proyectos especiales en función del Sexo

Antes de proceder a realizar la prueba de chi cuadrado se van a observar las distintas frecuencias observadas, para evitar que alguna sea demasiado pequeña.

```r
table(data$SpecialProjectsCount, data$Sex)
```

```
##
##       F   M
##   0 142  99
##   2   1   0
##   3   1   1
##   4   6   3
##   5  10  11
##   6  11  10
##   7   3   9
##   8   2   0
```

Como se puede ver, hay algunas celdas que son muy pequeñas como por ejemplo, los hombres que participane en 3 proyectos. De esta forma, se agruparán para evitar este problema. Se harán intervalos con 0 proyectos, de 1 a 5 y de 6 a 8.

```r
data[data$SpecialProjectsCount==0,'SpecialProjectsDisc'] <- 'No participa'
data[data$SpecialProjectsCount >= 1 & data$SpecialProjectsCount <= 5,'SpecialProjectsDisc'] <- 'Partici
data[data$SpecialProjectsCount>5,'SpecialProjectsDisc'] <- 'Participa en muchos'

table(data$SpecialProjectsDisc, data$Sex)
```

```
##
##                          F   M
##   No participa         142  99
##   Participa en algunos  18  15
##   Participa en muchos   16  19
```

Ahora si hay más cantidad en donde comprarar.

Hipótesis:

H_0 : La participación en proyectos especiales es independiente del sexo
H_1 : La participación en proyectos especiales si depende del sexo

Nivel de confianza: 95%; $\alpha = 0.05$

```r
chisq.test(data$SpecialProjectsDisc, data$Sex)
```

```
##
##  Pearson's Chi-squared test
##
## data:  data$SpecialProjectsDisc and data$Sex
## X-squared = 2.2621, df = 2, p-value = 0.3227
```

Como se observa, no existen diferencias significativas en las distribuciones de la variable participación en proyectos especiales en función del Sexo, lo que indica que no están relacionadas.

Además se considerará la misma clasificación de participación que en el sexo: no participación, participación y participación alta o en muchos proyectos.

Participación en proyectos especiales en función de la Raza

Hipótesis:

H_0 : La participación en proyectos especiales es independiente de ser blanco
H_1 : La participación en proyectos especiales si depende de ser blanco

Nivel de confianza: 95%; $\alpha = 0.05$

```r
chisq.test(data$SpecialProjectsDisc, data$RaceDescDisc)
```

```
##
##  Pearson's Chi-squared test
##
## data:  data$SpecialProjectsDisc and data$RaceDescDisc
## X-squared = 2.9909, df = 2, p-value = 0.2241
```

Como se observa en este caso tampoco hay discriminacion en función de si la persona es blanca o no. Dado que no depende serlo para tener mayor participación en proyectos especiales.

Estos modelos son relevantes porque muestran que no se les da más oportunidades de participar en proyectos importantes a determinadas personas por su sexo o raza.

Modelos de Regresión Lineal

Antes de proceder a realizar los modelos de regresión lineal se comprobará la relación de las variabes de salida con los posible predictores que se incluirán en los modelos. De esta forma se evitará meter información no relevante en las ecuaciones de regresión.

Las relaciones que ya se han comprobado son:
- Salario frente a Sexo y Raza
- Participación en proyectos importantes frente a Sexo y Raza

Las relaciones que quedarían por comprobar son:
- Salario frente a Departamento y Posición
- Participación en proyectos importantes frente a Departamento y Posición
- Salario frente a participación en proyectos importantes

Salario vs Departamento

Hipótesis:

H_0 : El salario es independiente del departamento
H_1 : El salario si depende del departamento

Nivel de confianza: 95%; $\alpha = 0.05$

```
kruskal.test(data$PayRate, data$Department)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  data$PayRate and data$Department
## Kruskal-Wallis chi-squared = 165.25, df = 4, p-value < 2.2e-16
```

Se ve claramente que el salario tiene una relación significativa con el salario.

Salario vs Posición

```
table(data$Position)
```

```
##
##                 Accountant I     Administrative Assistant
##                            3                            3
##           Area Sales Manager                 BI Developer
##                           27                            4
##                  BI Director                          CIO
##                            1                            1
##                 Data Analyst              Data Architect
##                            8                            1
##       Database Administrator       Director of Operations
##                            5                            1
##            Director of Sales         Enterprise Architect
##                            1                            1
##                  IT Director             IT Manager - DB
##                            1                            2
##           IT Manager - Infra         IT Manager - Support
```

```
##                            1                            1
##                   IT Support             Network Engineer
##                            8                            5
##              President & CEO     Principal Data Architect
##                            0                            1
##           Production Manager      Production Technician I
##                           14                          136
##     Production Technician II                Sales Manager
##                           57                            3
##          Senior BI Developer      Shared Services Manager
##                            3                            2
##            Software Engineer Software Engineering Manager
##                            9                            1
##               Sr. Accountant                       Sr. DBA
##                            2                            2
##         Sr. Network Engineer
##                            5
```

Como se puede ver en esta tabla, hay muchas posiciones y pocas personas en algunas de ellas. Es por eso que se agruparán las posiciones por niveles jerárquicos, distinguiendo así de profesionales, managers y directores.

```
professionals <- c('Accountant I', 'Administrative Assistant', 'BI Developer', 'Data Analyst', 'Data Ana
managers <- c('Area Sales Manager', 'IT Manager - DB', 'IT Manager - Infra', 'IT Manager - Support', 'Pr
directors <- c('BI Director', 'CIO', 'Director of Operations', 'Director of Sales', 'IT Director')

data$PositionReduced <- factor(data$Position, levels = c(professionals, managers, directors), labels = c
```

Ahora que las posiciones está agrupadas se realizará un test no paramétrico de Kruskal-Wallis, dado que hay más de 2 categorías en la variable independiente.

Hipótesis:

H_0 : El salario es independiente del nivel de responsabilidad
H_1 : El salario si depende del nivel de responsabilidad

Nivel de confianza: 95%; $\alpha = 0.05$

```
kruskal.test(data$PayRate, data$PositionReduced)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  data$PayRate and data$PositionReduced
## Kruskal-Wallis chi-squared = 110, df = 2, p-value < 2.2e-16
```

Se observa claramente que existe una relación entre el nivel de responsabilidad y el salario.

Se eliminarán las variables que carezcan de utilidad:

```
rm(directors, managers, professionals, emptydata)
```

Participación en proyectos importantes vs Departamento

Hipótesis:

H_0 : La participación en proyectos importantes es independiente del departamento
H_1 : La participación en proyectos importantes si depende del departamento

Nivel de confianza: 95%; $\alpha = 0.05$

```
chisq.test(data$SpecialProjectsDisc, data$Department)
```

```
## Warning in chisq.test(data$SpecialProjectsDisc, data$Department): Chi-squared
## approximation may be incorrect

##
##  Pearson's Chi-squared test
##
## data:  data$SpecialProjectsDisc and data$Department
## X-squared = 349.66, df = 8, p-value < 2.2e-16
```

Sale una advertencia porque las cantidades que hay en cada nivel de participación y departamento es muy pequeño, tal y como se puede comprobar en la siguiente tabla:

```
table(data$SpecialProjectsCount, data$Department)
```

```
##
##     Admin Offices IT/IS Production Sales Software Engineering
##   0             2     0        208    31                    0
##   2             1     0          0     0                    0
##   3             1     0          0     0                    1
##   4             3     2          0     0                    4
##   5             2    16          0     0                    3
##   6             1    18          0     0                    2
##   7             0    12          0     0                    0
##   8             0     2          0     0                    0
```

Es por eso que el resultado de la prueba se tomará con precaución a la hora de aplicarlo en los modelos de regresión.

Participación en proyectos importantes vs Posición

```
table(data$SpecialProjectsDisc, data$PositionReduced)
```

```
##
##                         prof managers dtor
##   No participa           193       46    2
##   Participa en algunos    30        2    1
##   Participa en muchos     30        3    2
```

Se obseva que hay unos tamaños en la tabla de contingencia muy pequeños distintos de 0. En este sentido se aplicará el test de Fisher, que tiene la misma función que el test de Chi-Cuadrado, pero para muestras muy pequeñas.

Hipótesis:

H_0 : La participación en proyectos importantes es independiente de la posición
H_1 : La participación en proyectos importantes si depende de la posición

Nivel de confianza: 95%; $\alpha = 0.05$

```r
fisher.test(data$SpecialProjectsDisc, data$PositionReduced)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  data$SpecialProjectsDisc and data$PositionReduced
## p-value = 0.03183
## alternative hypothesis: two.sided
```

Tal y como se muestra, la probabilidad asociada es menor que el nivel de significación, por lo que se rechaza la hipótesis nula y se considera algún tipo de relación entre el nivel de participación en proyectos importantes y el de responsabilidad.

Salario vs Participación en proyectos importantes

Por último se verá si existe una relación entre la participación en proyectos importantes y el salario.

Hipótesis:

H_0 : El salario es independiente de la participación en proyectos importantes
H_1 : El salario si depende de la participación en proyectos importantes

Nivel de confianza: 95%; $\alpha = 0.05$

```r
kruskal.test(data$PayRate, data$SpecialProjectsDisc)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  data$PayRate and data$SpecialProjectsDisc
## Kruskal-Wallis chi-squared = 64.58, df = 2, p-value = 9.476e-15
```

La probabilidad es menor que el nivel de significación, por lo que se rechaza la hipótesis nula de independencia de las variables y se considera que hay una relación entre el nivel de participación en proyectos importantes y el salario.

Una vez vistas las relaciones, se podría pensar que el salario vendría determinado por el sexo, el departamento y el nivel de participación en proyectos importantes y este último, vendría determinado por el nivel de responsabilidad.

En este sentido se van a probar 2 modelos de regresión lineal:
- El salario en función del sexo, el departamento, el nivel de participación en proyectos importantes y el nivel de responsabilidad
- El salario en función del sexo, el departamento y el nivel de participación en proyectos importantes

Así se podrá comprobar si el nivel de responsabilidad aporta al modelo algo más que su efecto conjunto con el nivel de participación en proyectos importantes.

Modelo de Regresión Lineal para el salario

Se construirá un modelo de regresión lineal del salario en función de: el sexo, el departamento, la participación en proyectos importantes y el nivel de responsabilidad.

```r
set.seed(123)
train_select <- sample(seq(1,length(data$PayRate),1), size = 0.8 * length(data$PayRate))
train_data <- data[train_select,]
test_data <- data[-train_select,]
```

```r
modelo <- lm(PayRate ~ Sex + Department + SpecialProjectsDisc + PositionReduced, data = train_data)

summary(modelo)
```

```
##
## Call:
## lm(formula = PayRate ~ Sex + Department + SpecialProjectsDisc +
##     PositionReduced, data = train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.530  -2.980   0.026   3.216  18.053
##
## Coefficients:
##                                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                               34.22108    7.20811   4.748 3.56e-06
## SexM                                      -0.80414    0.95289  -0.844   0.3996
## DepartmentIT/IS                           14.77718    3.26717   4.523 9.64e-06
## DepartmentProduction                     -12.43678    7.17520  -1.733   0.0843
## DepartmentSales                           -0.02632    7.10469  -0.004   0.9970
## DepartmentSoftware Engineering            20.48121    3.68709   5.555 7.44e-08
## SpecialProjectsDiscParticipa en algunos   -8.82391    7.75785  -1.137   0.2565
## SpecialProjectsDiscParticipa en muchos    -4.24696    7.93012  -0.536   0.5928
## PositionReducedmanagers                   21.58305    1.75184  12.320  < 2e-16
## PositionReduceddtor                       28.08988    4.14332   6.780 9.50e-11
##
## (Intercept)                              ***
## SexM
## DepartmentIT/IS                          ***
## DepartmentProduction                     .
## DepartmentSales
## DepartmentSoftware Engineering           ***
## SpecialProjectsDiscParticipa en algunos
## SpecialProjectsDiscParticipa en muchos
## PositionReducedmanagers                  ***
## PositionReduceddtor                      ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.962 on 237 degrees of freedom
## Multiple R-squared:  0.8046, Adjusted R-squared:  0.7972
## F-statistic: 108.4 on 9 and 237 DF,  p-value: < 2.2e-16
```

```r
confusion_matrix <- cbind(test_data$PayRate, predict(modelo, test_data))
sqrt(mean((confusion_matrix[,1] - confusion_matrix[,2])^2))
```

```
## [1] 4.622478
```

Como se observa, anteriormente se vio que el sexo mostraba influencia sobre el salario, pero se diluye su efecto junto a otras variables.

Por otro lado, pertenecer al dpto. de IT o Software Engeenering y ser director o manager es un factor que favorece un salario más elevado. Mientras que pertenecer al departamento de Production es un factor que favorece un salario más bajo (factor negativo).

Este modelo explica el 79% de la varianza del salario, por lo que es un modelo bastante bueno.

Además, una vez aplicado al conjunto de test, se ve puede extraer un error medio de 4.62, lo que significa que el modelo tiene una precisión bastante buena cuando estima el salario por hora.

Ahora se buscará el modelo de regresión lineal para el salario en función del departamento y la participación en proyectos importantes. Se obvia el sexo dado que no tuvo efecto significativo en el anterior modelo.

```
modelo <- lm(PayRate ~ Department + SpecialProjectsDisc, data = train_data)

summary(modelo)
```

```
##
## Call:
## lm(formula = PayRate ~ Department + SpecialProjectsDisc, data = train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.355  -4.480  -1.274   1.735  36.726
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           55.0000     9.1447   6.014 6.69e-09
## DepartmentIT/IS                       16.6863     4.2374   3.938 0.000108
## DepartmentProduction                 -31.7258     9.1727  -3.459 0.000642
## DepartmentSales                        0.6019     9.3125   0.065 0.948524
## DepartmentSoftware Engineering        22.5649     4.8221   4.680 4.81e-06
## SpecialProjectsDiscParticipa en algunos -29.7418     9.8878  -3.008 0.002910
## SpecialProjectsDiscParticipa en muchos  -24.3308    10.1350  -2.401 0.017127
##
## (Intercept)                          ***
## DepartmentIT/IS                      ***
## DepartmentProduction                 ***
## DepartmentSales
## DepartmentSoftware Engineering       ***
## SpecialProjectsDiscParticipa en algunos **
## SpecialProjectsDiscParticipa en muchos  *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.145 on 240 degrees of freedom
## Multiple R-squared:  0.6586, Adjusted R-squared:  0.6501
## F-statistic: 77.17 on 6 and 240 DF,  p-value: < 2.2e-16
```

```
confusion_matrix <- cbind(test_data$PayRate, predict(modelo, test_data))
sqrt(mean((confusion_matrix[,1] - confusion_matrix[,2])^2))
```

```
## [1] 7.361655
```

En este caso el modelo aporta menos explicación de la varianza que el anterior. Además aporta un error cuadrático medio mayor que el anterior, por lo que se estaría cometiendo un error mayor utilizando este modelo.

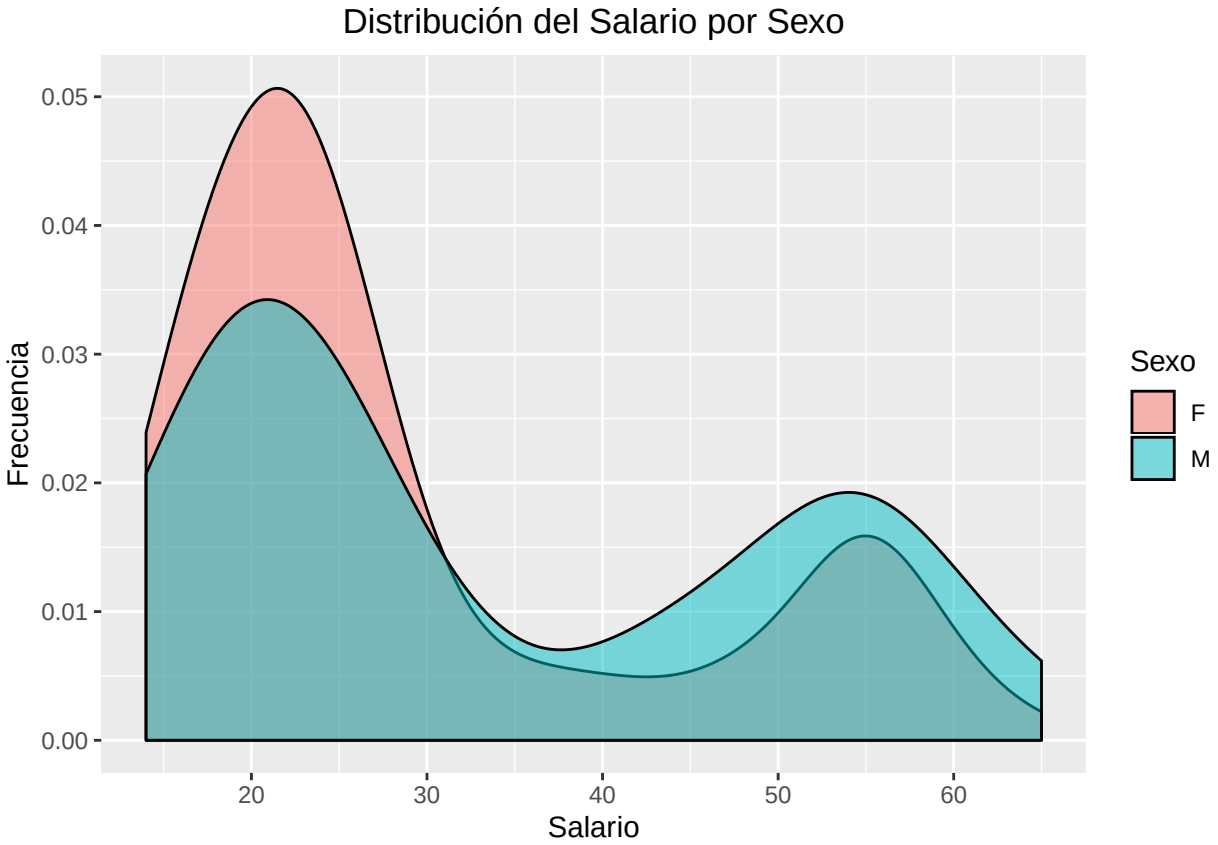
Se considerará por tanto el anterior modelo dado que se entiende que la variable nivel de responsabilidad aporta un efecto al salario diferente del que aporta junto con el nivel de participación en proyectos importantes.

5. REPRESENTACIÓN DE RESULTADOS A PARTIR DE GRÁFICAS Y TABLAS

En la siguiente tabla se pueden ver las diferentes pruebas realizadas y la descripción de la conclusión.

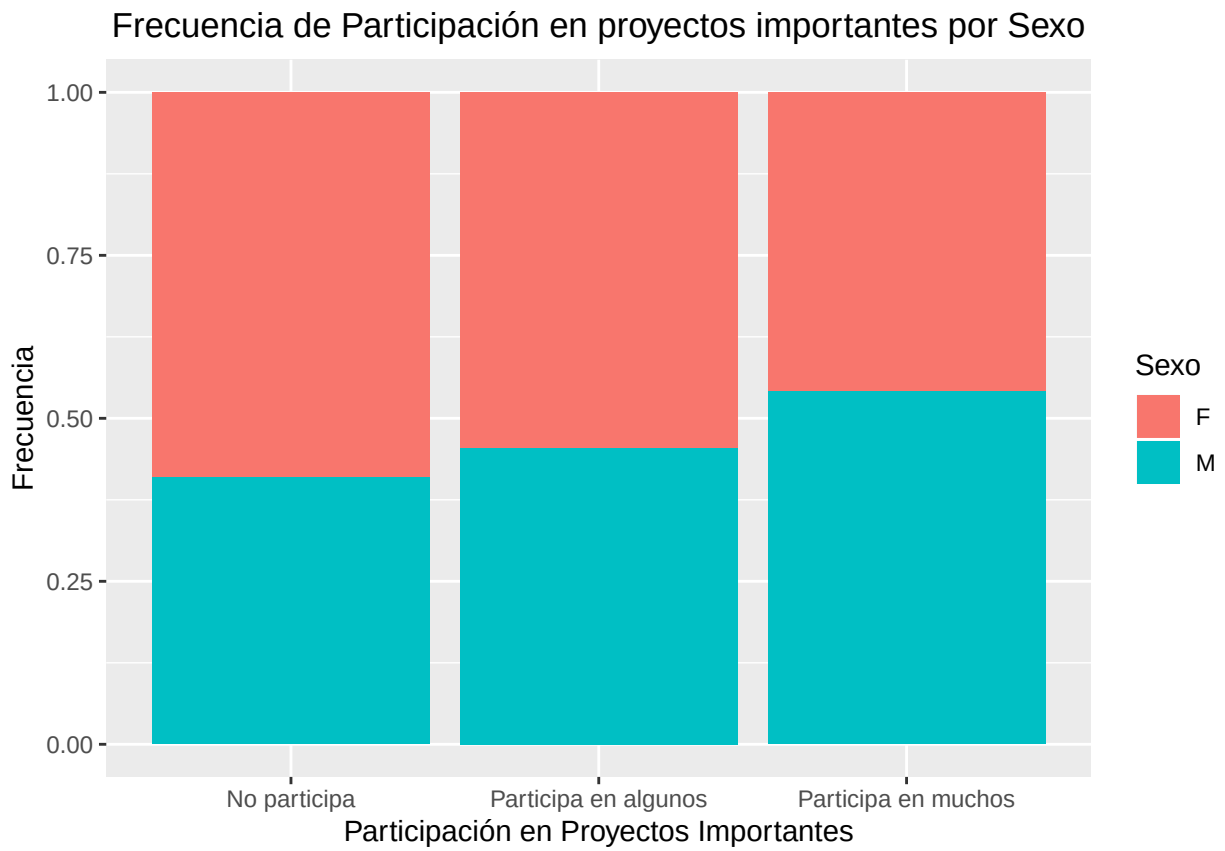
VARIABLES	TEST	HIPÓTESIS_0	CONCLUSIÓN
Salario - Sexo	Mann-Whitney	rechazada	Existe efecto significativo del sexo sobre el salario
Proyectos - Sexo	Chi-Squared	aceptada	No hay efecto significativo del sexo sobre la participación en proyectos importantes
Salario - Raza	Kruskal-Wallis	aceptada	No hay efecto significativo de la raza sobre el salario
Proyectos - Raza	Chi-Squared	aceptada	No hay efecto significativo de la raza sobre la participación en proyectos importantes
Salario - Dpto	Kruskal-Wallis	rechazada	Existe efecto significativo del sexo sobre el salario
Proyectos - Dtpo	Chi-Squared	?	No hay conclusión en base a la prueba
Salario - Posición	Kruskal-Wallis	rechazada	Existe efecto significativo del nivel de responsabilidad sobre el salario
Proyectos - Posición	Kruskal-Wallis	rechazada	Existe efecto significativo del nivel de resposnabilidad sobre el de participación en proyectos importantes
Salario - Proyectos	Kruskal-Wallis	rechazada	Existe efecto significativo del nivel de participación en proyectos importantes sobre el salario

```
ggplot(data, aes(PayRate, fill = Sex)) + geom_density(alpha = 0.5) + labs(title = 'Distribución del Sala
```



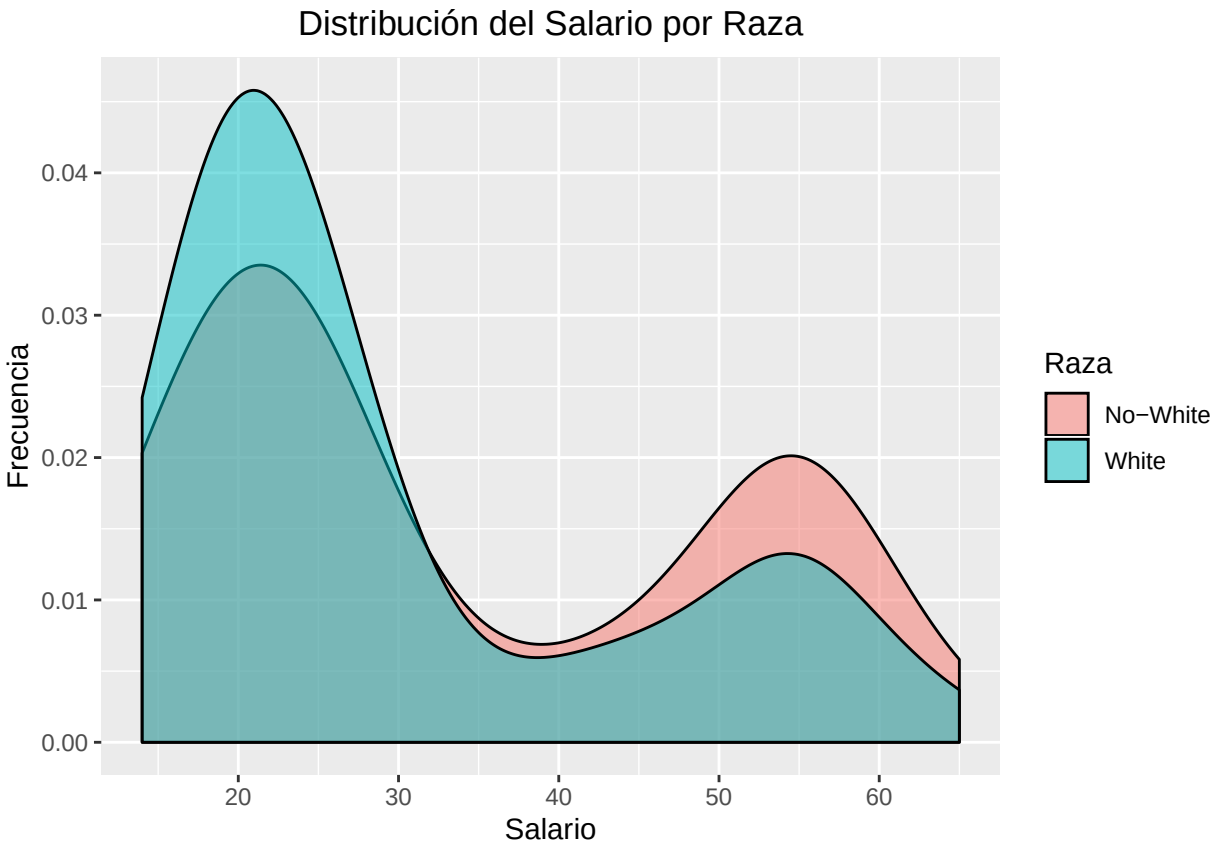
Se observa que a pesar de que la prueba mostrase una relación entre las variables, el efecto que posteriormente se diluyó entre otras variables tiene una explicación: la diferencia de salarios por sexo no es apenas significativa.

```
ggplot(data, aes(SpecialProjectsDisc, fill = Sex)) + geom_bar(alpha = 1, position = 'fill') + labs(titl
```



Se ve que la participaciónen proyectos no depende mucho del sexo, algo que va acorde con la conclusión de la prueba.

```
ggplot(data, aes(PayRate, fill = RaceDescDisc)) + geom_density(alpha = 0.5) + labs(title = 'Distribució
```



En esta relación pasa algo parecido a la relación del sexo con el salario. No hay mucha diferencia entre las distribuciones en función de la raza.

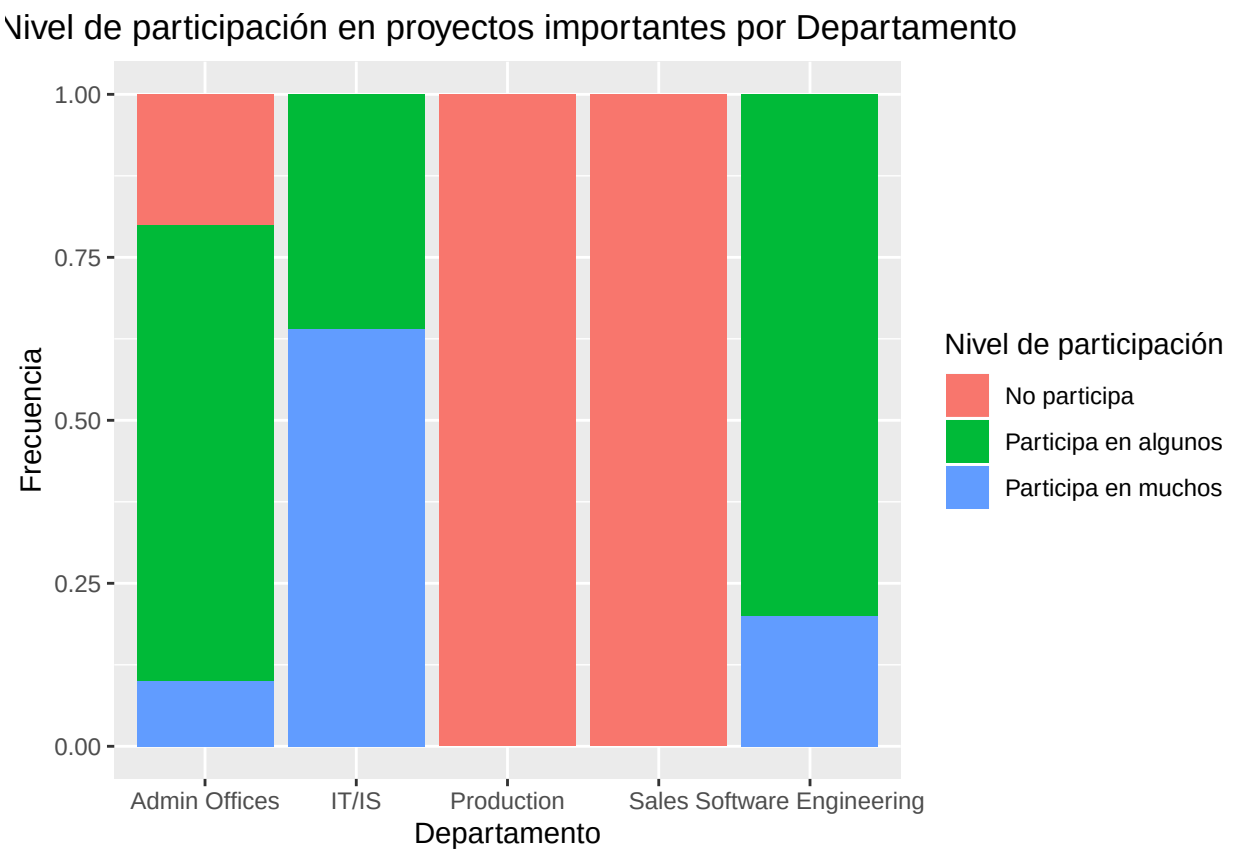
```
ggplot(data, aes(SpecialProjectsDisc, fill = RaceDescDisc)) + geom_bar(alpha = 1, position = 'fill') + l
```

Como se observa, la conclusión de la prueba es acertada dado que hay aproximadamente la misma distribución en cada nivel de participación en proyectos importantes en función de la raza.

```
ggplot(data, aes(PayRate, x=Department)) + geom_boxplot(alpha = 1, fill = 'steelblue2') + labs(title =
```

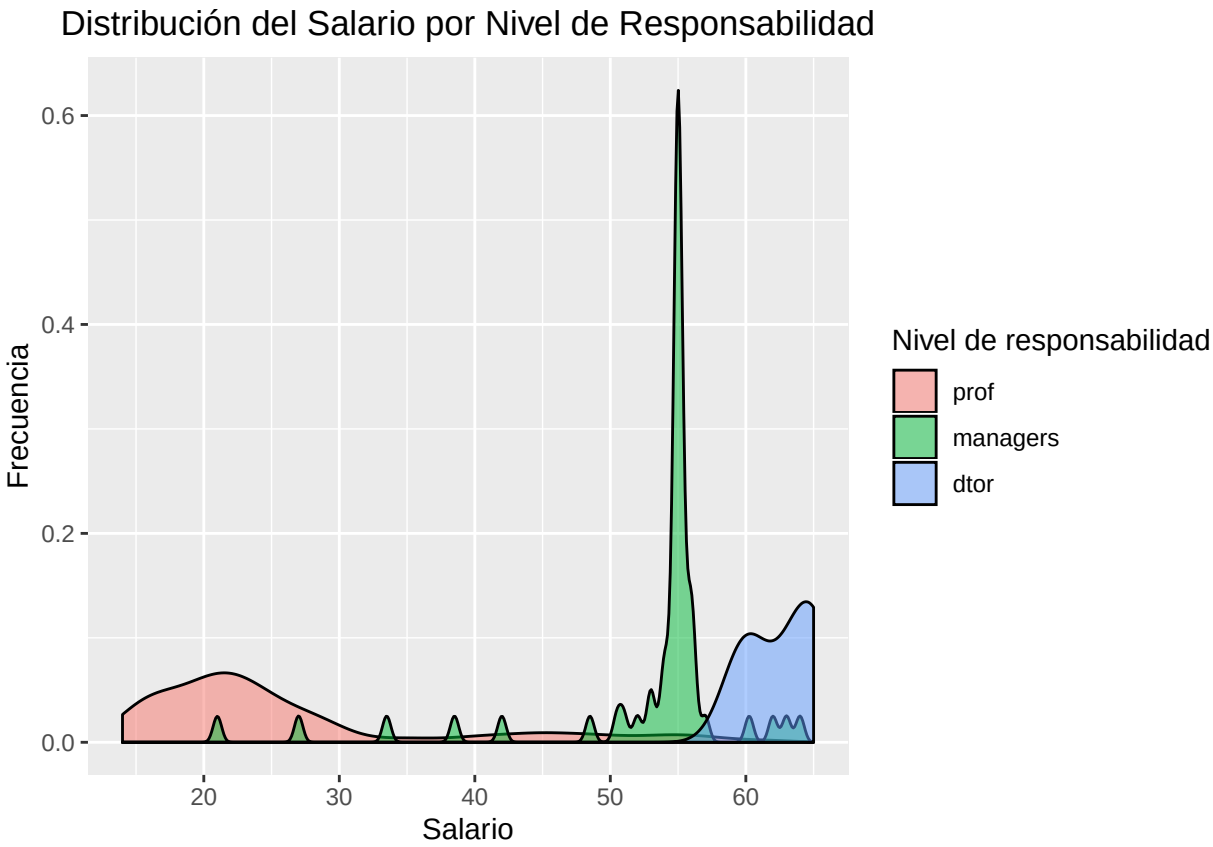
Como se observa en el gráfico, los diferentes departamentos tienen bastante diferencia entre sí, y no son tan dispersos dentro de los mismos.

```
ggplot(data, aes(fill=SpecialProjectsDisc, Department)) + geom_bar(alpha = 1, position = 'fill') + labs
```



Se ve que hay diferencias en las distribuciones en el nivel de participación en proyectos importantes a lo largo de los diferentes departamentos. Por ejemplo, así como empleados de Production y Sales no participan, los de Software Engineering tienen al menos 2 proyectos importantes.

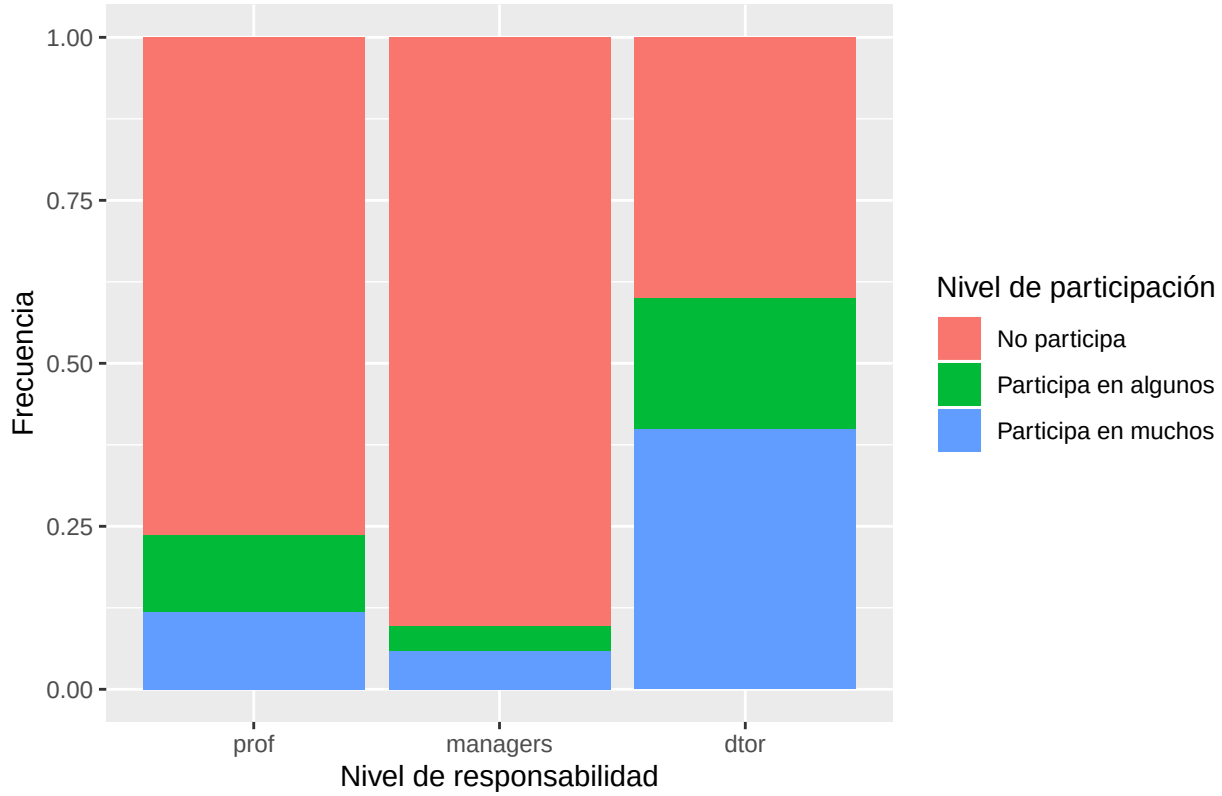
```
ggplot(data, aes(PayRate, fill = PositionReduced)) + geom_density(alpha = 0.5) + labs(title = 'Distribu
```



En el gráfico se observa que en función del nivel de responsabilidad se podrían separar 3 distribuciones del salario, lo cual indica que hay una cierta relación entre estas 2 variables.

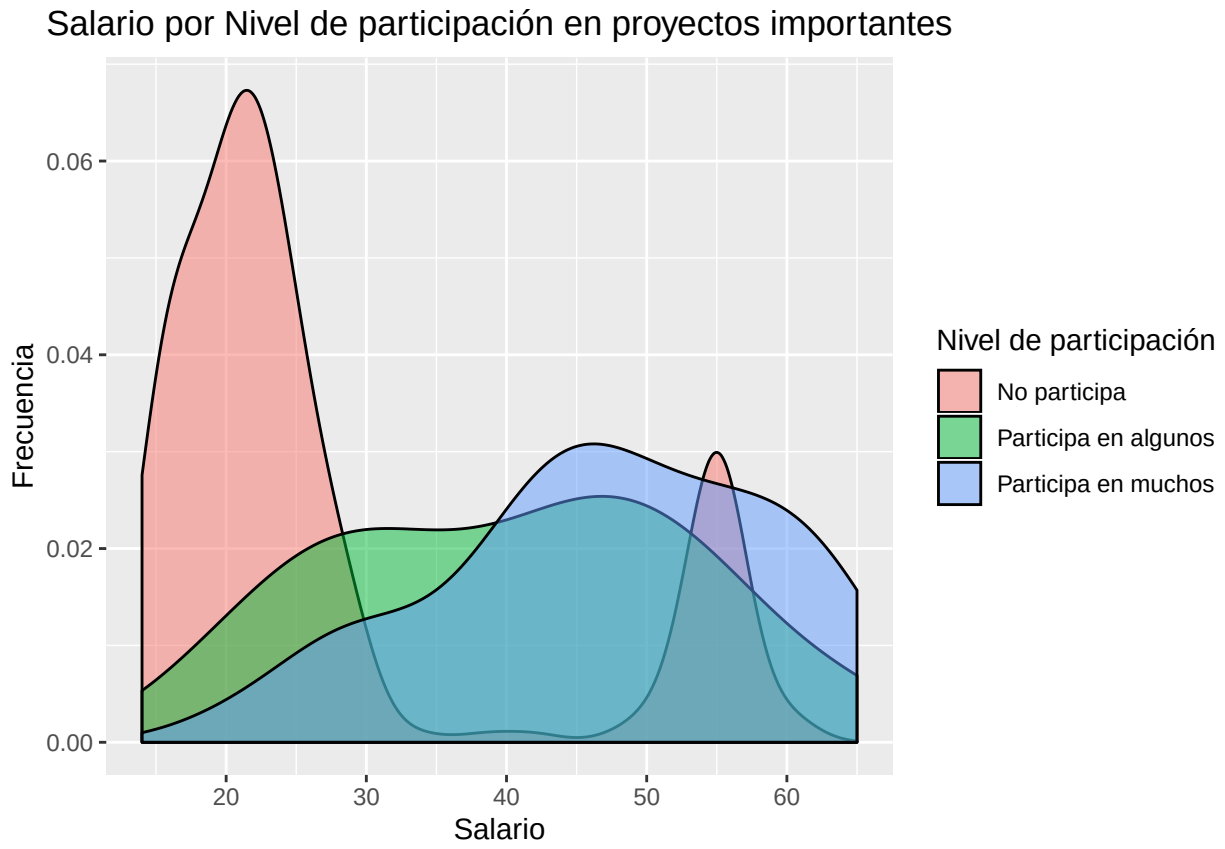
```
ggplot(data, aes(fill = SpecialProjectsDisc, PositionReduced)) + geom_bar(alpha = 1, position = 'fill')
```

de Participación en proyectos importantes por Nivel de responsabilidad



También se puede observar una cierta diferencia en las distribuciones del nivel de participación a lo largo del nivel de responsabilidad.

```
ggplot(data, aes(PayRate, fill = SpecialProjectsDisc)) + geom_density(alpha = 0.5) + labs(title = 'Sala
```



Se puede notar una cierta diferencia en los 3 niveles de participación en proyectos importantes si se observan los extremos. Así como es mu fácil encontrar salarios más bajos en los que no participan en ningún proyecto importante, es bastante más fácil encontrar salarios muy altos si se participa en muchos.

En cuanto a los modelos de regresión:

MODELO	RECTA REG.	R^2	RMSE
Modelo 1	$S = 34.22 + 14.78IT + 20.48SE + 21.58M + 28.09D + 6.96$	0.8	4.62
Modelo 2	$S = 55 + 16.69IT + 22.56SE - 29.74PM - 24.33PA + 9.15$	0.65	7.36

Donde S = Salario, IT = pertenece al dpto. IT, SE = Pertenece al dpto. de Software Engineering, M = es manager, D = es director, PM = nivel medio de participación en proyectos importantes y PA = alto nivel de participación en proyectos importantes.

6. RESOLUCIÓN DEL PROBLEMA

Las conclusiones a las que se llegan son las siguientes:

- La empresa es capaz de controlar el sesgo de género y raza a la hora de establecer las compensaciones económicas de sus empleados.
- La empresa es capaz, además, de controlar estos sesgos también a la hora de ofrecer oportunidades de desarrollo a los empleados, entendiendo las oportunidades de desarrollo como la participación en proyectos de mayor impacto para la empresa.

- La empresa parece tener una estrategia de negocio en la que se diferencian los tipos de perfiles y departamentos clave, lo que se refleja en la relación entre las oportunidades de desarrollo y compensación con la posición y el área del empleado. Esto además se hace de manera bastante homogénea, dado que es posible estimar con bastante acierto la compensación económica de los empleados en función del perfil y el área.

CONTRIBUCIONES	FIRMA
Investigación previa	RSM
Redacción de las respuestas	RSM
Desarrollo de código	RSM