

Pdf de experimento para no equivocarme en el importante

Nazareth Artigas C.I: 31.375.207 Santiago López C.I: 30.330.479

2026-02-21

2. Planteamiento del Problema

En la última década, el mercado laboral global ha experimentado una transformación sin precedentes. Factores como la digitalización acelerada, la normalización del trabajo remoto y la demanda de habilidades técnicas específicas han redefinido cómo se valoran los puestos de trabajo. Ya no basta con mirar el “título del cargo”; la compensación económica ahora parece estar sujeta a una red compleja de variables interconectadas.

¿Es el tamaño de la empresa un predictor fiable del sueldo? ¿Existe una penalización salarial por elegir la modalidad remota, o es ahora un estándar competitivo? ¿Qué tanto peso real tienen ciertas habilidades (skills) frente a la ubicación geográfica?. A pesar de la abundancia de ofertas de trabajo, existe una falta de claridad sobre cuáles son los verdaderos determinantes del salario en el mercado actual, por lo cual, este estudio busca aclarar esas dudas mediante la estadística descriptiva y el uso de la plataforma GitHub y lenguajes de programación conocidos como R y Python.

3. Objetivos

3.1 Objetivo General

Analizar el panorama laboral de los sectores tecnológicos emergentes mediante estadística descriptiva para identificar las variables que más impactan en la oferta salarial y la demanda de talento.

3.2 Objetivos Específicos

- Determinar el promedio, mediana y desviación estándar de los salarios por industria para identificar cuál es la más rentable.
- Comparar las medias salariales entre puestos remotos y presenciales para verificar si existe una “penalización” o beneficio económico por la modalidad.
- Identificar las ciudades con mayores ofertas de empleo y su nivel de remuneración promedio.
- Clasificar las habilidades más solicitadas en cada sector y su frecuencia de aparición en las vacantes.
- Evaluar si el tamaño de la empresa (Small, Medium, Large) influye significativamente en los rangos salariales ofrecidos.

#Limpieza de datos

```
# 1. Cargar la librería silenciosamente  
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --  
## v dplyr     1.2.0    v readr     2.2.0  
## vforcats   1.0.1    v stringr   1.6.0  
## v ggplot2   4.0.2    v tibble    3.3.1  
## v lubridate 1.9.5    v tidyr    1.3.2
```

```

## v purrr      1.2.1
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
# 2. Leer el archivo usando comillas dobles
# Asegúrate de que el nombre sea EXACTO al que ves en tu carpeta de archivos
datos <- read_csv("10. Job y Skills.xlsx")

## Multiple files in zip: reading 'xl/drawings/drawing1.xml'
## Rows: 1 Columns: 1
## -- Column specification -----
## Delimiter: ","
## chr (1): <?xml version="1.0" encoding="UTF-8" standalone="yes"?>
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
# 3. Limpiar los nombres de las columnas automáticamente
# (Esto quita espacios y puntos raros en los títulos de las columnas)
install.packages("janitor") # Ejecuta esto una vez si no la tienes

## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.5'
## (as 'lib' is unspecified)

library(janitor)

##
## Attaching package: 'janitor'
##
## The following objects are masked from 'package:stats':
##       chisq.test, fisher.test
datos <- datos %>% clean_names()

# 4. Revisar si funcionó
glimpse(datos)

## Rows: 1
## Columns: 1
## $ xml_version_1_0_encoding_utf_8_standalone_yes <chr> "<xdr:wsDr xmlns:xdr=\\"h~
# 1. Cargamos la librería para Excel
library(readxl)

# 2. Intentamos leer el archivo (usa el nombre exacto que aparece en tu panel de Files)
# Si le cambiaste el nombre a "dataset.xlsx", pon ese aquí.
datos <- read_excel("10. Job y Skills.xlsx")

# 3. Revisamos qué hay adentro
head(datos)

## # A tibble: 6 x 9
##   job_id job_title     industry location salary_usd skills_required remote_option
##   <dbl> <chr>        <chr>     <chr>          <dbl> <chr>           <chr>
## 1      1 Quantum Res~ Quantum~ Singapo~     175780 Linear Algebra~ No

```

```

## 2      2 Renewable E~ Green T~ Singapo~      137481 Climate Data A~ Yes
## 3      3 Quantum Res~ Quantum~ Tokyo      182081 Linear Algebra~ No
## 4      4 Sustainabil~ Green T~ Singapo~      113822 Climate Data A~ No
## 5      5 Smart Contr~ Blockch~ London      92575 Rust, Solidity Yes
## 6      6 Smart Contr~ Blockch~ Tokyo      173379 Solidity, Rust Yes
## # i 2 more variables: company_size <chr>, posting_date <dttm>
# Esto nos mostrará la estructura de las columnas y los primeros valores
glimpse(datos)

## Rows: 10,000
## Columns: 9
## $ job_id          <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, ~
## $ job_title        <chr> "Quantum Researcher", "Renewable Energy Engineer", "Qu~
## $ industry         <chr> "Quantum Computing", "Green Tech", "Quantum Computing"~
## $ location          <chr> "Singapore", "Singapore", "Tokyo", "Singapore", "Londo~
## $ salary_usd       <dbl> 175780, 137481, 182081, 113822, 92575, 173379, 99659, ~
## $ skills_required   <chr> "Linear Algebra, Quantum Algorithms", "Climate Data An~
## $ remote_option     <chr> "No", "Yes", "No", "No", "Yes", "Yes", "No", "Yes", "N~
## $ company_size      <chr> "Large", "Large", "Medium", "Large", "Small", "Medium"~
## $ posting_date       <dttm> 2025-07-22, 2025-09-26, 2025-12-31, 2025-05-29, 2025~~

# 1. Ver cuántos valores faltantes hay en total
sum(is.na(datos))

## [1] 0

# 2. Ver cuáles columnas tienen esos valores vacíos
colSums(is.na(datos))

##      job_id      job_title      industry      location      salary_usd
##            0            0            0            0            0
## skills_required remote_option company_size posting_date
##            0            0            0            0

```