

Análisis del mercado laboral en España

Catret Ruber, Pablo¹, Palazón Caballero, José Miguel^{1,*}, Rosique Martínez, Marcos¹

¹ Universitat de València - Escola Tècnica Superior d'Enginyeria (ETSE) Avinguda de l'Universitat, 46100 Burjassot, Valencia;

* Correspondence: jomipaca@alumni.uv.es.

Simple Summary: A Simple summary goes here.

Abstract: A single paragraph of about 200 words maximum. For research articles, abstracts should give a pertinent overview of the work. We strongly encourage authors to use the following style of structured abstracts, but without headings: 1) Background: Place the question addressed in a broad context and highlight the purpose of the study; 2) Methods: Describe briefly the main methods or treatments applied; 3) Results: Summarize the article's main findings; and 4) Conclusion: Indicate the main conclusions or interpretations. The abstract should be an objective representation of the article, it must not contain results which are not presented and substantiated in the main text and should not exaggerate the main conclusions.

Keywords: keyword 1; keyword 2; keyword 3 (list three to ten pertinent keywords specific to the article, yet reasonably common within the subject discipline.).

1. Introducción

El mercado laboral en España es un sistema complejo y dinámico que refleja las interacciones entre diversos factores económicos, sociales y demográficos. Este trabajo de análisis exploratorio de datos combina dos pilares fundamentales para entender las dinámicas laborales: la Encuesta de Población Activa (EPA) y la Clasificación Nacional de Actividades Económicas (CNAE-2009). A través de estas fuentes, se busca ofrecer una visión integral de las tasas de empleo, actividad y ocupación, desglosadas por género, grupo de edad y comunidad autónoma, así como por ramas de actividad económica.

La EPA, realizada trimestralmente por el Instituto Nacional de Estadística (INE), proporciona información detallada sobre la participación laboral de la población, permitiendo analizar las diferencias en el acceso al empleo según el género, la edad y el territorio. En esta parte del estudio, se examinan las tasas del mercado laboral en las distintas comunidades autónomas y cómo estas se ven influenciadas por eventos como la crisis financiera de 2008 o la pandemia de COVID-19. Además, se busca identificar desigualdades estructurales y dinámicas regionales que puedan servir como base para estudios más específicos o el diseño de políticas públicas.

Por otro lado, la CNAE-2009 aporta un marco estándar para clasificar las actividades económicas en las que se desempeñan los trabajadores, permitiendo analizar cómo se distribuye la fuerza laboral entre diferentes sectores. Este enfoque complementario permite explorar no solo el "dónde" y "quién" trabaja, sino también el "en qué" trabaja la población, revelando patrones de especialización sectorial y el impacto de la transformación económica a lo largo de los años.

La combinación de ambas perspectivas —la distribución demográfica y regional desde la EPA y la estructura sectorial desde la CNAE— permite abordar preguntas clave sobre el mercado laboral, tales como:

- ¿Cómo varía la participación laboral entre comunidades autónomas?
- ¿Cómo varía la participación laboral entre grupos de edad?

Citation: Catret Ruber, P.; Palazón Caballero, J.M.; Rosique Martínez, M. Análisis del mercado laboral en España. *Journal Not Specified* **2023**, *1*, 0. <https://doi.org/>

Received:

Revised:

Accepted:

Published:

Copyright: © 2025 by the authors. Submitted to *Journal Not Specified* for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

- ¿Qué sectores han experimentado los mayores cambios tras eventos históricos como la crisis de 2008 o la pandemia de COVID-19?

Este análisis exploratorio busca no solo describir el estado actual del mercado laboral en España, sino también identificar tendencias, desigualdades y oportunidades de mejora. El objetivo final es proporcionar una base sólida para futuros estudios y contribuir al diseño de estrategias efectivas en el ámbito económico, social y político.

2. Importación y tratamiento de datos

En primer lugar, se han descargado los datos directamente desde la base de datos abierta INE Base. Estos datos se presentan en un formato de csv, separado por el carácter ‘;’, y con el uso de marca de decimales española ‘,’.

2.1. Encuesta de Población Activa (EPA)

Como se ha mencionado anteriormente en la introducción, la EPA es realizada trimestralmente por el Instituto Nacional de Estadística, por lo que se han extraído cinco datasets de su base de datos: población total, activa, inactiva, ocupada y parada.

Cabe destacar que estos datasets están desglosados por comunidad y ciudad autónoma, sexo y grupo de edad, con las tres columnas expresadas en “miles de personas” como unidad. El periodo de los datos figura desde el primer trimestre de 2002 hasta el tercer trimestre de 2024.

```
Poblacion <- read_delim("data/Poblacion Total.csv",
  delim = ";", escape_double = FALSE, col_types = cols(Total = col_number()),
  locale = locale(decimal_mark = ",", grouping_mark = "."),
  trim_ws = TRUE)

Activos <- read_delim("data/Activos.csv",
  delim = ";", escape_double = FALSE, col_types = cols(Total = col_number()),
  locale = locale(decimal_mark = ",", grouping_mark = "."),
  trim_ws = TRUE)
Activos <- Activos %>% rename(`Activos` = `Total`)

Inactivos <- read_delim("data/Inactivos.csv",
  delim = ";", escape_double = FALSE, col_types = cols(Total = col_number()),
  locale = locale(decimal_mark = ",", grouping_mark = "."),
  trim_ws = TRUE)
Inactivos <- Inactivos %>% rename(`Inactivos` = `Total`)

Ocupados <- read_delim("data/Ocupados.csv",
  delim = ";", escape_double = FALSE, col_types = cols(Total = col_number()),
  locale = locale(decimal_mark = ",", grouping_mark = "."),
  trim_ws = TRUE)
Ocupados <- Ocupados %>% rename(`Ocupados` = `Total`)

Parados <- read_delim("data/Parados.csv",
  delim = ";", escape_double = FALSE, col_types = cols(Total = col_number()),
  locale = locale(decimal_mark = ",", grouping_mark = "."),
  trim_ws = TRUE)
Parados <- Parados %>% rename(`Parados` = `Total`)
```

Al tratarse de un análisis del mercado laboral eliminaremos al grupo de población menor a 16 años pues no tienen edad suficiente para trabajar. Notamos que en el dataset Población hay una columna que solo contiene la cadena “Total Nacional”, por lo que la eliminamos. Además, en la columna “Comunidades y Ciudades Autónomas” hay

valores faltantes que coinciden con las filas del propio Total Nacional, de modo que los completamos con dicha cadena.

```
Poblacion_edad_trabajar <- Poblacion %>%
  rename(`Población en edad de trabajar` = `Total`) %>%
  filter(Edad != "Menores de 16") %>%
  select(-contains("Total Nacional")) %>%
  mutate(`Comunidades y Ciudades Autónomas` = replace_na(`Comunidades y Ciudades A
```

Notamos que para el conjunto de datos de población total en edad de trabajar y población inactiva hay un intervalo de edad más que para el resto de datasets: se divide l"55 y más años" en los grupos "De 55 a 64 años" y "65 y más años". Dado que buscamos obtener un dataset único y compacto uniremos ambos grupos de edad como en el resto de datasets.

```
Poblacion_edad_trabajar <- Poblacion_edad_trabajar %>%
  mutate(Edad = ifelse(Edad %in% c("De 55 a 64 años", "65 y más años"), "55 y más
group_by(Sexo, `Comunidades y Ciudades Autónomas`, Edad, Periodo) %>%
  summarise(`Población en edad de trabajar` = sum(`Población en edad de trabajar`,

Inactivos <- Inactivos %>%
  mutate(Edad = ifelse(Edad %in% c("De 55 a 64 años", "65 y más años"), "55 y más
group_by(Sexo, `Comunidades y Ciudades Autónomas`, Edad, Periodo) %>%
  summarise(`Inactivos` = sum(`Inactivos`, na.rm = TRUE), .groups = 'drop')
```

Una vez hemos normalizado la estructura de los datasets podemos unificar todos los datos en uno solo.

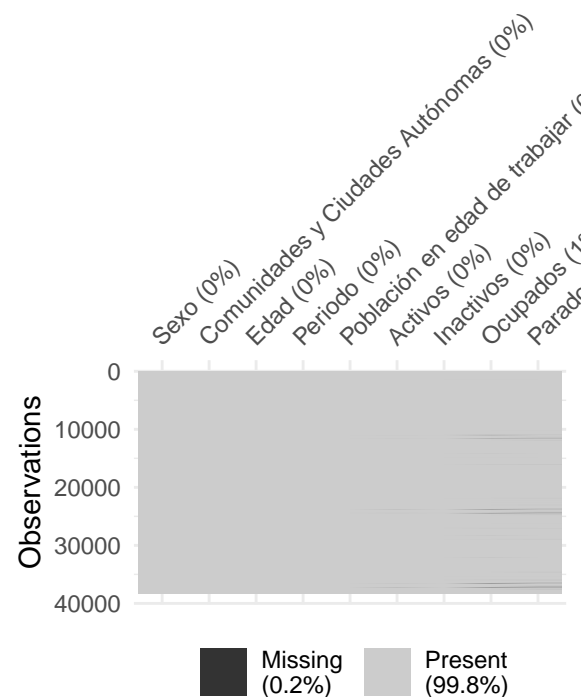
```
datos_EPA <- left_join(Poblacion_edad_trabajar, Activos, by=c("Sexo", "Comunidades
left_join(., Inactivos, by=c("Sexo", "Comunidades y Ciudades Autón
left_join(., Ocupados, by=c("Sexo", "Comunidades y Ciudades Autóno
left_join(., Parados, by=c("Sexo", "Comunidades y Ciudades Autónom
```

Observamos que los datos de la columna "Periodo" son cadenas; para solucionarlo, definimos y empleamos la función "SacarFechas" para transformarlos a tipo Date. Del mismo modo

```
datos_EPA$Periodo <- as.Date(sapply(datos_EPA$Periodo, sacarFechas), format = "%Y-
# Borramos los datasets innecesarios
rm(list = c("Poblacion", "Poblacion_edad_trabajar", "Activos", "Inactivos", "Ocupa
```

Observemos si existen datos faltantes en nuestro dataset.

```
vis_miss(datos_EPA)
```



Existe únicamente un porcentaje muy bajo de valores faltantes en la columna “Parados”. Veamos un ejemplo de estos casos.

```
miss_datos_EPA <- datos_EPA %>% mutate(miss_parados = is.na(Parados)) %>% filter(miss_parados)
library(tibble)
library(dplyr)
library(ggplot2)
library(readr)
library(readxl)
library(rstan)
library(rstanarm)
library(shinystan)
library(tibble)
library(tidyverse)
library(viridis)

## $'Comunidades y Ciudades Autónomas'
##
##           02 Aragón           03 Asturias, Principado de
##                1                      3
##           06 Cantabria 15 Navarra, Comunidad Foral de
##                11                      11
##           16 País Vasco           17 Rioja, La
##                2                      14
##           18 Ceuta           19 Melilla
##                79                      151
##
## $Edad
##
## 55 y más años De 16 a 19 años De 20 a 24 años De 25 a 34 años De 35 a 44 años
##           101           114           15           1
## De 45 a 54 años           Total
##           27           1
```

Los datos se concentran en Comunidades y Ciudades Autónomas con proporcionalmente poca población y en rangos de edades donde es poco habitual estar parado (ya que para que se cumpla dicha condición se debe estar buscando activamente trabajo). De hecho, el único valor que podría resultar extraño es que exista un valor faltante para un Total de edades, pero al comprobar la localización y fecha notamos que es razonable.

```
miss_datos_EPA[miss_datos_EPA$Edad == "Total",c("Sexo", "Comunidades y Ciudades Au

## # A tibble: 1 x 3
##   Sexo   'Comunidades y Ciudades Autónomas' Período
##   <chr>   <chr>                                <date>
## 1 Hombres 19 Melilla                        2002-06-01

rm(list=c("miss_datos_EPA"))
```

Dicho patrón parece reflejar que muy poca población de esas características se encuentra parada; por tanto, sustituiremos los valores faltantes del dataset por ceros.

```
datos_EPA <- datos_EPA %>% mutate(Parados = ifelse(is.na(Parados), 0, Parados))
```

2.2. Clasificación Nacional de Actividades Económicas (CNAE-2009)

Este dataset divide los datos según rama de actividad, sexo y fecha en el periodo comprendido entre el primer trimestre de 2008 y el tercer trimestre de 2024, usando las mismas unidades que en el resto de datasets. No obstante, en este caso podemos encontrar los datos segmentado en dos subconjuntos principales:

- Porcentajes ("Total_abs"): Representan la proporción de ocupación de cada rama de actividad con respecto al total del sexo correspondiente.
- Valores absolutos ("Total_porc"): Proporcionan el número total de empleados en cada rama.

Esta separación permite analizar tanto las tendencias globales (valores absolutos) como la estructura relativa de los sectores (porcentajes).

```
datos_CNAE <- read.csv("data/CNAE.csv", sep=";", dec=",")
```

Esta vez al usar la función "read.csv" tendremos que modificar el tipo de dato de la columna "Total" de cadena de texto a numérico. Además, nos encontramos de nuevo con el problema del formato de las fechas, por lo que volveremos a aplicar la función SacarFechas.

```
datos_CNAE$Total <- as.numeric(gsub(",", ".", gsub("\\.", "", datos_CNAE$Total)))
datos_CNAE$Periodo <- as.Date(sapply(datos_CNAE$Periodo, sacarFechas), format = "%
```

Para diferenciar entre cada tipo de dato (valor absoluto y porcentaje) vamos a extraerlos en dos datasets para posteriormente realizar un join, esencialmente como si se realizara un pivote.

```
datos_CNAE_por <- datos_CNAE[datos_CNAE$Unidad=="Porcentaje",]
datos_CNAE_por <- datos_CNAE_por[, !names(datos_CNAE_por) %in% "Unidad"]

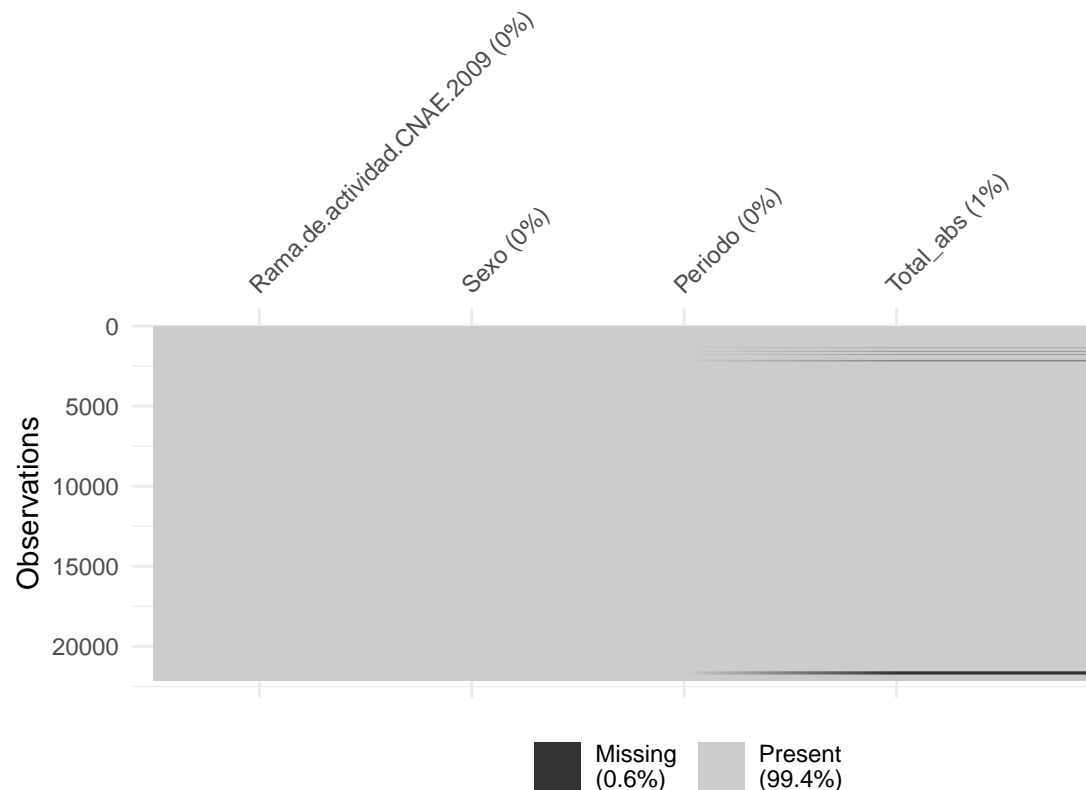
datos_CNAE_abs <- datos_CNAE[datos_CNAE$Unidad=="Valor absoluto",]
datos_CNAE_abs <- datos_CNAE_abs[, !names(datos_CNAE_abs) %in% "Unidad"]

if (dim(datos_CNAE_abs)[1] == dim(datos_CNAE_por)[1] & dim(datos_CNAE_abs)[2] ==
    datos_CNAE <- left_join(datos_CNAE_abs, datos_CNAE_por, by = c("Rama.de.actividad"))
}

rm(list = c("datos_CNAE_abs", "datos_CNAE_por"))
```

Una vez tenemos la tabla unificada es recomendable realizar un análisis de datos faltantes.

```
vis_miss(datos_CNAE)
```



Observamos que para un pequeño subconjunto de filas no existen cifras totales. Veamos si siguen algún patrón.

```
miss_datos_CNAE <- datos_CNAE %>% mutate(miss_total = is.na(Total_abs)) %>% filter(
  head(unique(miss_datos_CNAE$Rama.de.actividad.CNAE.2009), 4)
```

```
## [1] "05 Extracción de antracita, hulla y lignito"
## [2] "06 Extracción de crudo de petróleo y gas natural"
## [3] "07 Extracción de minerales metálicos"
## [4] "09 Actividades de apoyo a las industrias extractivas"
```

```
rm(list=c("miss_datos_CNAE"))
```

Notamos que los valores faltantes se encuentran en ramas de actividades poco comunes relacionadas con la industria pesada, por lo que dichos NA reflejarán que muy poca población se dedica a ello, probablemente menos del mínimo registrable; en consecuencia sustituiremos dichos valores por ceros.

```
datos_CNAE <- datos_CNAE %>% mutate(Total_abs = ifelse(is.na(Total_abs), 0, Total_
```

Por último, añadiremos una columna que calcula la variación entre un periodo y el siguiente, lo que permitirá identificar momentos de cambio significativo en el mercado laboral, detectando tendencias positivas o negativas a lo largo del tiempo. Esto también permitirá ver de manera rápida la estacionalidad de la ocupación, sobre todo en ciertas ramas de actividad.

Es importante mencionar que realizar las diferencias entre un periodo y el siguiente genera NAs para el primer periodo de cada rama, por lo que sustituiremos estos valores faltantes por ceros.

```

datos_CNAE$Diferencia_Total <- NA

for (sexo in unique(datos_CNAE$Sexo)) {
  for (rama in unique(datos_CNAE$Rama.de.actividad.CNAE.2009)) {
    subset <- datos_CNAE[datos_CNAE$Sexo == sexo &
                        datos_CNAE$Rama.de.actividad.CNAE.2009 == rama, ]

    datos_CNAE$Diferencia_Total[datos_CNAE$Sexo == sexo &
                                datos_CNAE$Rama.de.actividad.CNAE.2009 == rama & d
                                diff(subset$Total_abs)
  }
}

datos_CNAE$Diferencia_Total[datos_CNAE$Periodo == min(datos_CNAE$Periodo)] <- 0
rm(list = c("subset"))

```

Appendix C

Appendix C.1

The appendix is an optional section that can contain details and data supplemental to the main text. For example, explanations of experimental details that would disrupt the flow of the main text, but nonetheless remain crucial to understanding and reproducing the research shown; figures of replicates for experiments of which representative data is shown in the main text can be added here if brief, or as Supplementary data. Mathematical proofs of results not central to the paper can be added as an appendix.

Appendix D

All appendix sections must be cited in the main text. In the appendixes, Figures, Tables, etc. should be labeled starting with 'A', e.g., Figure A1, Figure A2, etc.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.