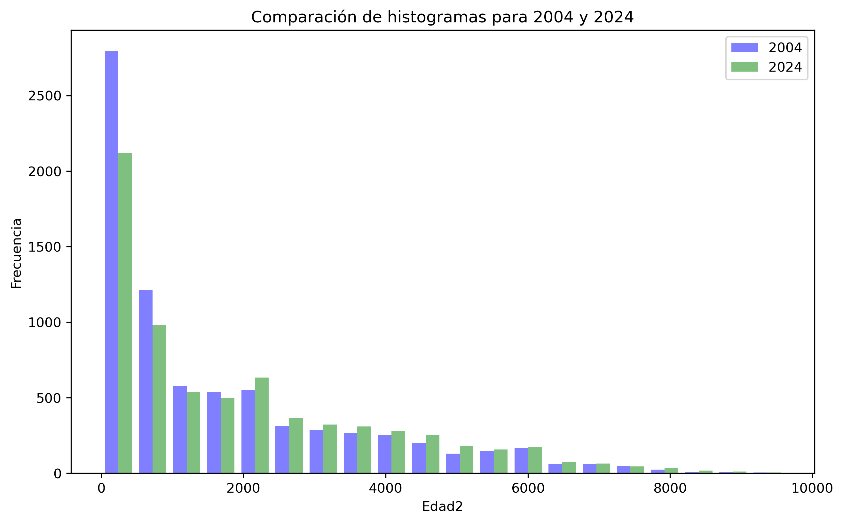
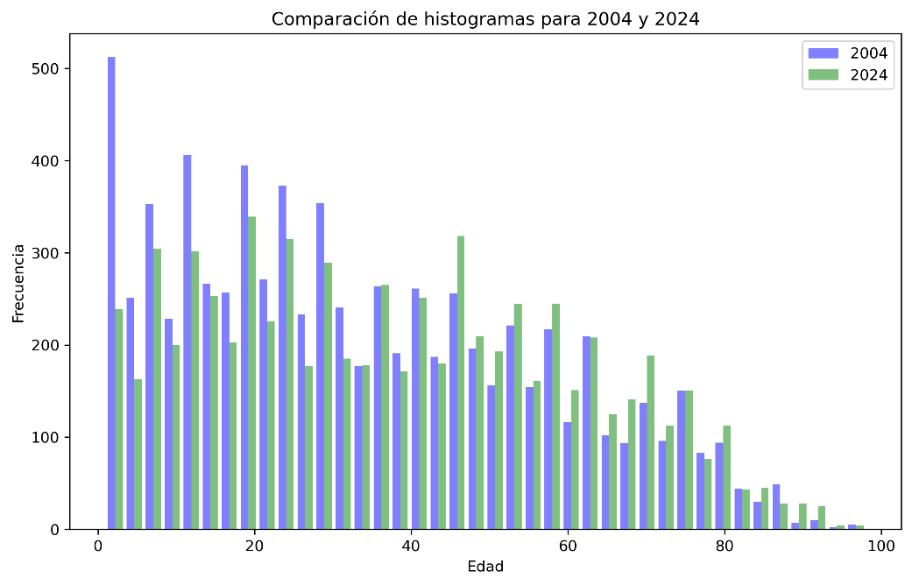
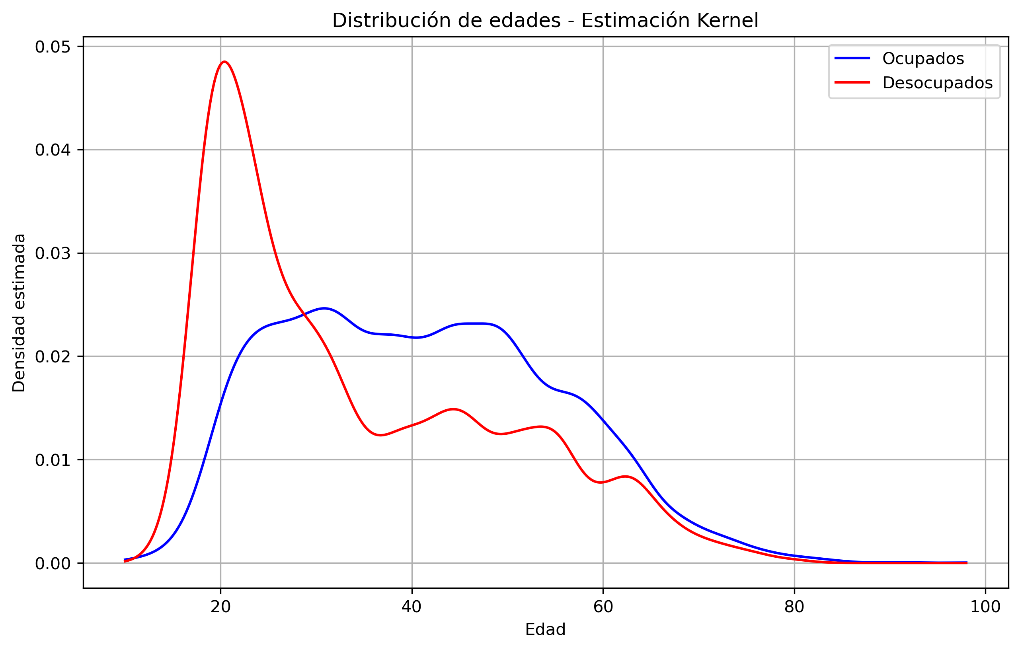
TP3 Big data and Machine Learning  
Franco Bustelo  
Juan Pablo Tomasello  
David Agudelo.

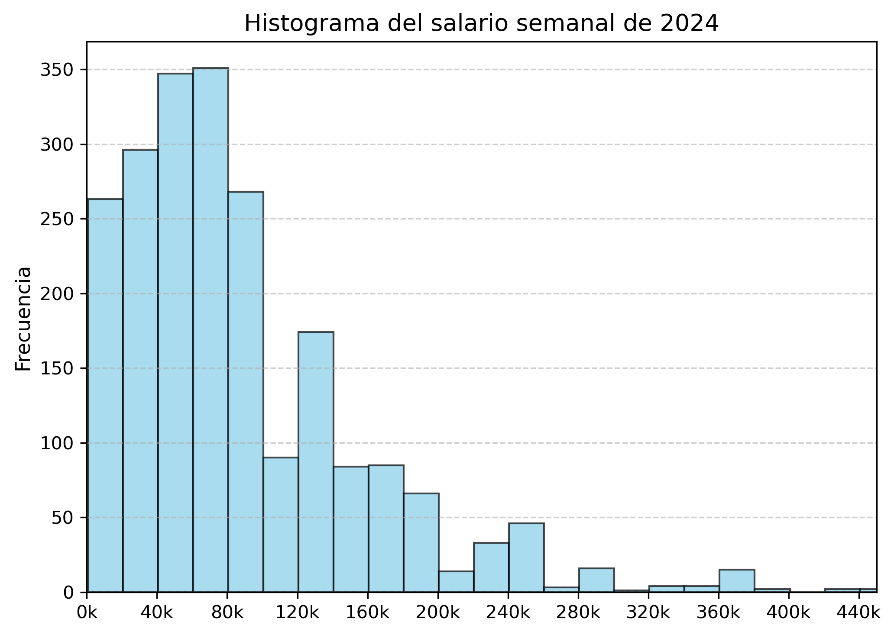
El gráfico muestra un histograma comparativo de la variable **Edad2** (elevada al cuadrado y luego con raíz cuadrada, por lo que es equivalente a Edad2) para los años 2004 (en azul) y 2024 (en verde).

El gráfico compara la distribución de la **edad al cuadrado** entre 2004 y 2024. En ambos años, la mayor frecuencia se da en valores bajos, lo que indica predominancia de personas jóvenes. Sin embargo, en 2024 hay un leve aumento en los valores más altos, lo que sugiere un envejecimiento relativo de la población en ese período.

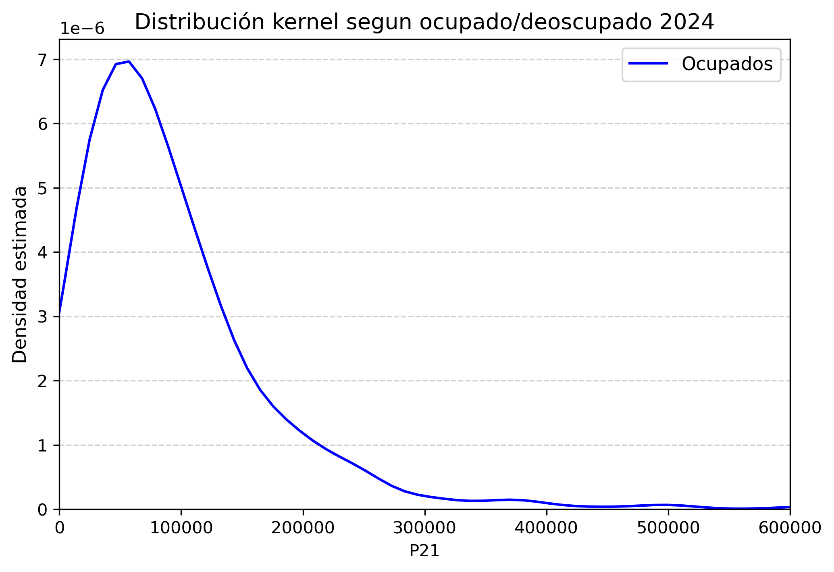


****El gráfico muestra la distribución de la **edad** entre 2004 y 2024. En 2004 predominan los grupos más jóvenes, especialmente menores de 30 años. En 2024 se observa una mayor frecuencia relativa en edades medias y altas (de 40 años aproximadamente), lo que indica un claro **envejecimiento de la población** en este periodo.

La mayor densidad de desocupados se concentra en edades **jóvenes de alrededor de 20 años,** mientras que los ocupados presentan una distribución **que abarca más edades**, con mayor densidad entre los **22 y 60 años aproximadamente**. Esto podría tratarse de un indicador de que el desempleo afecta principalmente a personas jóvenes.

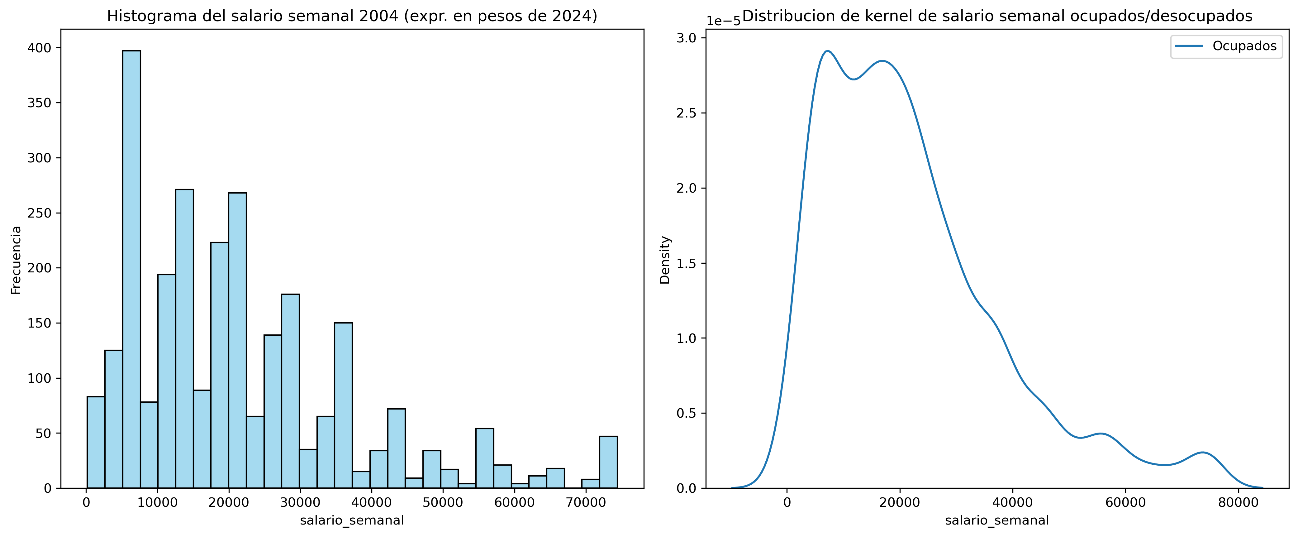


**Histograma del salario semanal**   
Muestra que la mayoría de los ocupados (done más frecuencia hay) ganan entre 20000 y 180000 pesos semanales. Teniendo valores más bajos en cuanto a frecuencia en salarios por encima de los 200000



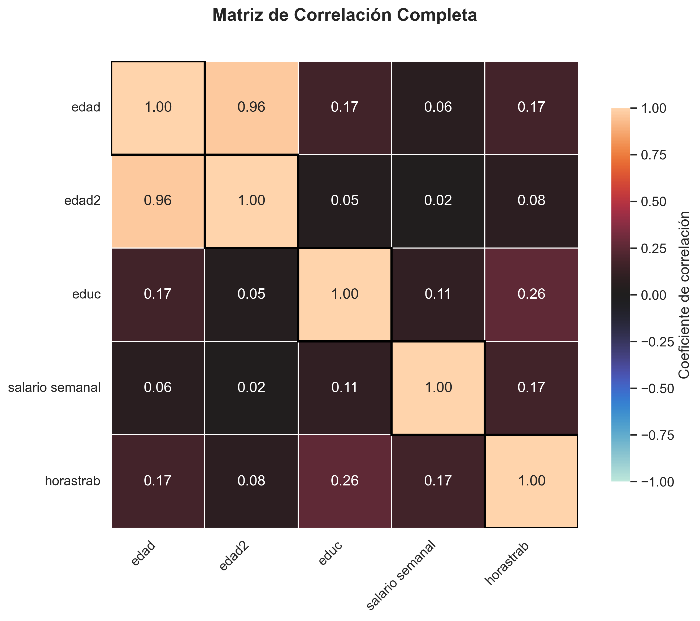
**Distribución kernel según ocupado o desocupado**   
Se ve una gran concentración hasta los salarios de 200000 y muestra, con una larga cola a la derecha los valores no tan recurrentes. Este histograma nos está indicando que hay pocos casos con ingresos altos.

En la consigna se pide hacerlo con desocupados también pero, como mostramos en los códigos, no hay ningún desocupado que tenga algún salario ya que todos son 0

En el primer panel se ve que el salario semanal de 2004 expresado en pesos de 2024 tiene su mayor concentración de frecuencia entre los 5000 y los 40000 pesos semanales. Contando con una mucha menor frecuencia por encima de dichos 40000 pesos.

En el otro panel se puede ver la distribución de kernel de dichos salarios semanales pero esta vez teniendo en cuenta el estado, es decir, si están ocupados o desocupados. Se puede ver como los ocupados rondan dentro del rango entre los mismos 5000 y los 40000 pesos una vez más. Esto quiere decir que la mayoría de ocupados tienen su salario semanal rondado por dichos valores. En la consigna también se pide tener en cuenta a los desocupados pero estos, como se explicó y demostró en los códigos no presentan ningún salario. Todo aquel que sea desocupado tiene un p21 de valor 0, lo cual tendría sentido ya que un desempleado no debería porque tener un salario semanal.

|  |  |  |  |
| --- | --- | --- | --- |
|  | 2004 | 2024 | Total |
| Cant. de observaciones | 7647 | 7051 | 14698 |
| Can. De observaciones con nas en la variable “Estado” | 10 | 41 | 51 |
| Cantidad de ocupados | 3079 | 3224 | 6303 |
| Cantidad de desocupados | 528 | 311 | 839 |
| Cantidad de variables limpias y homogeneizadas | 311 | 232 | 543 |

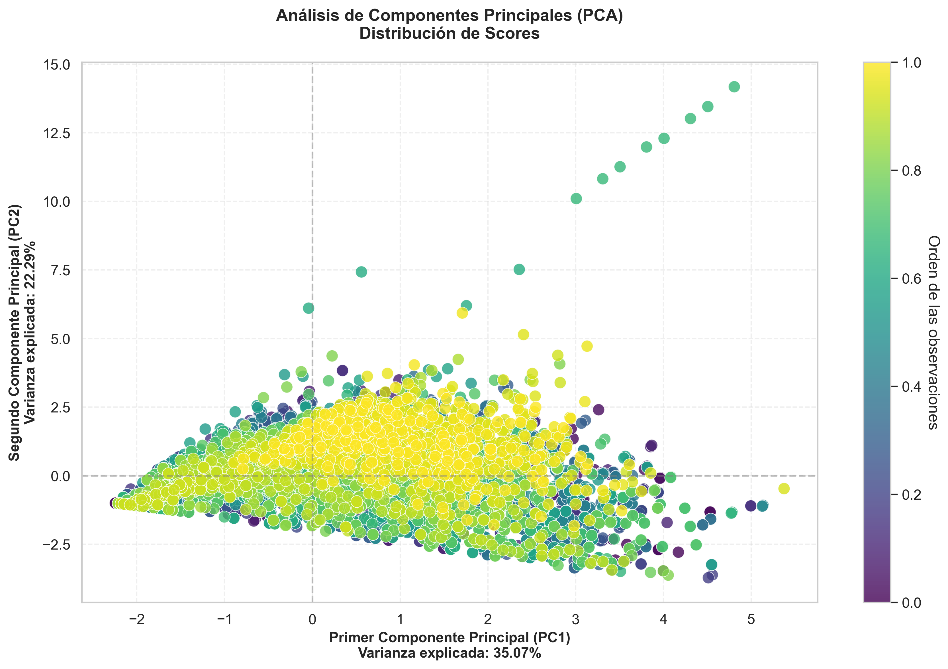


La matriz correlación nos muestra, obviando la correlación entre ‘edad’ y ‘edad2’ ( de **0.96** que es bastante obvia ya que son lo mismo pero elevada al cuadrado), que, la más interesante es la que se observa entre ‘educ’ y ‘horastrab’, con un valor de **0.26**. Esto quiere decir que las personas con mayor nivel educativo tienden, en promedio, a trabajar más horas.

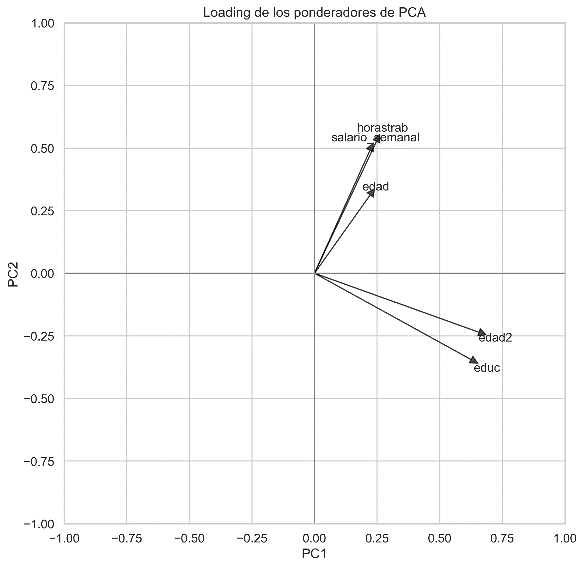
Por otro lado, la correlación entre ‘educ’ y ‘salario\_semanal’ es más baja (de un **0.11**), lo cual indica que al tener un mayor nivel educativo no necesariamente se traduce en un salario más alto.

Otra correlación que se tiene en cuenta es la de ‘salario\_semanal’ con ‘horastrab’ la cual es de **0.17**. Esto quiere decir que hay una relación muy baja entre la cantidad de horas que una persona trabaja y lo que gana por semana. Es decir, trabajar más horas no necesariamente significa tener un salario más alto.

Finalmente, se observa que ni ‘edad’ ni ‘edad2’ tienen una relación significativa con el salario semanal (entre **0.06** y **0.02** respectivamente) ni con las horas trabajadas (en torno a **0.17 y 0.08**). Es decir, la edad no parece ser un factor determinante ni en el ingreso semanal ni en la carga horaria laboral de las personas encuestadas.

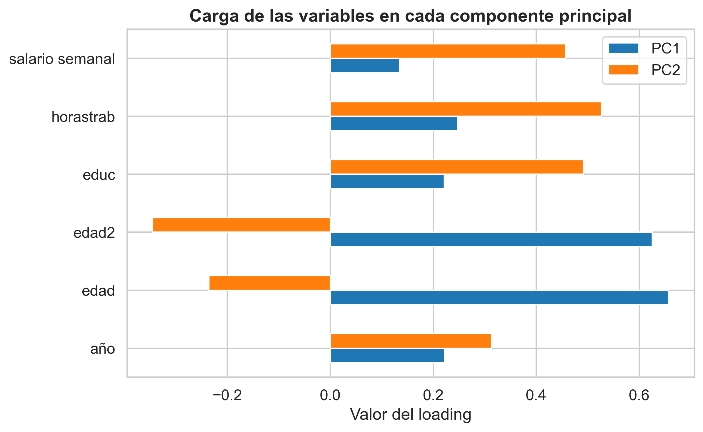
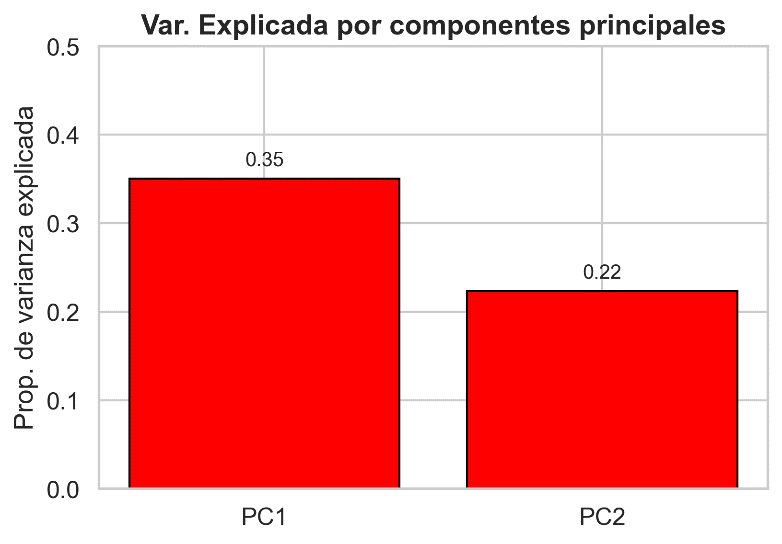


La interpretación ponela aca

  
**‘Educ’** y ‘**edad2’** están bien juntas y apuntan en la misma dirección, lo que significa que están muy relacionadas entre sí. Además, estas variables tienen una mayor carga sobre el eje **PC1**, o sea que influyen más en ese componente. Como ambas tienen que ver con el nivel educativo y la edad, parece que PC1 capta algo relacionado a eso, como el “perfil formativo” o el "nivel personal".

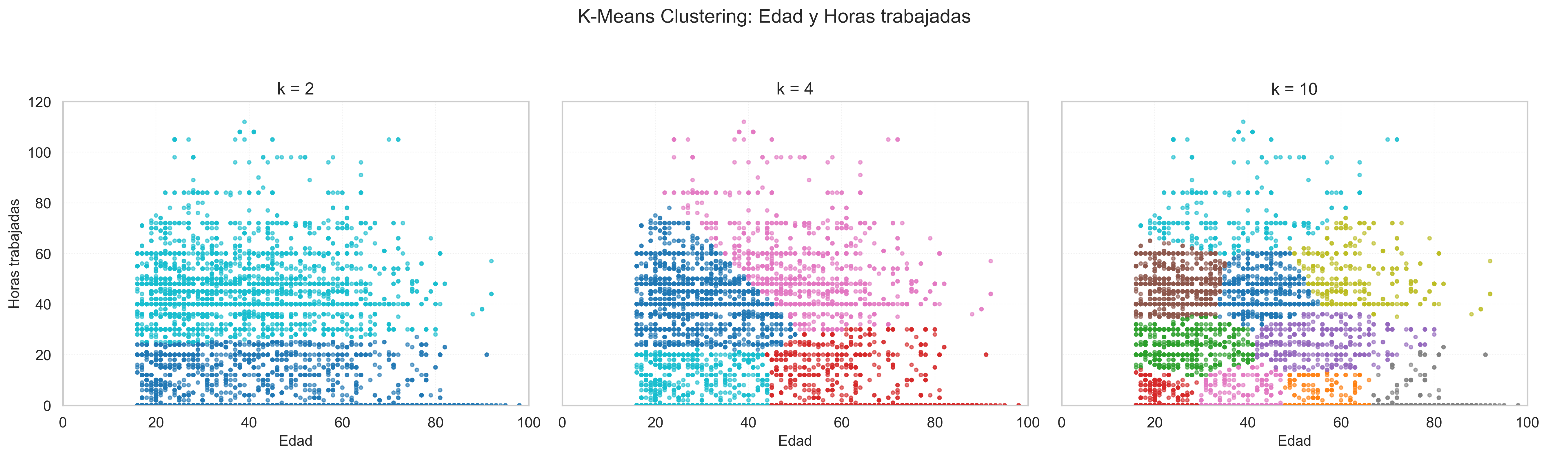
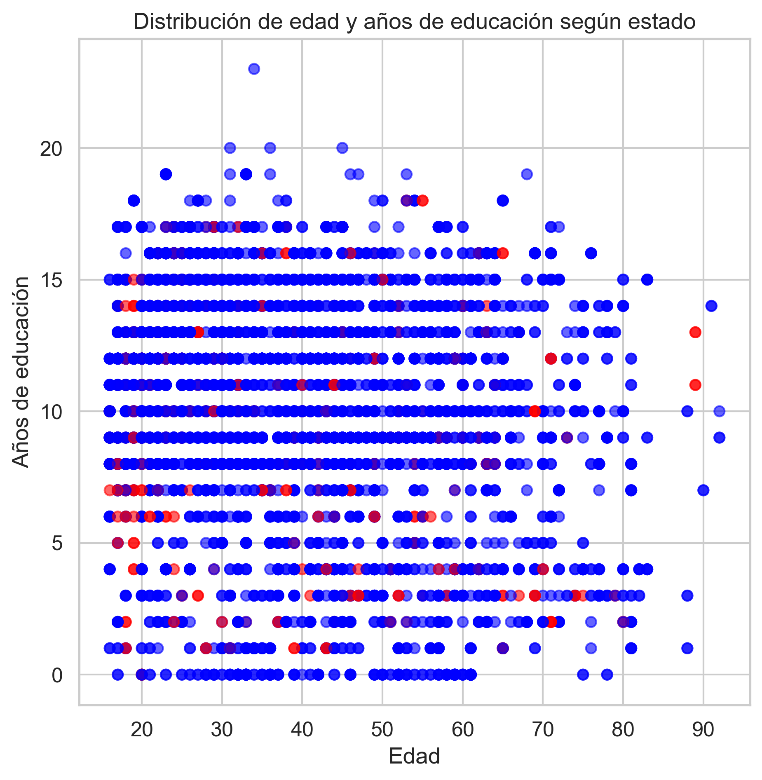
Por otro lado, **horastrab**, **salario\_semanal** y **edad** apuntan más hacia arriba (positivo para el PC2), lo que indica que tienen más peso en ese eje. Como todas están relacionadas al trabajo, es posible que **PC2** represente algo así como la responsabilidad con el trabajo.

También es interesante ver que las flechas no están todas en la misma dirección. Eso nos dice que las variables están captando cosas diferentes, lo cual puede ser bueno para el análisis porque ayuda a que los componentes principales separen bien la información.



En este gráfico se ve que el primer componente principal (PC1) está fuertemente influenciada por la edad, mientras que la segunda (PC2) representa más el aspecto laboral, con carga alta en salario semanal, educación y horas trabajadas. Esto muestra que las variables se agrupan en dos ejes principales: uno más personal y otro más laboral.  
  
  
  
  
  
  
PC1 explica el 35% de la varianza total del conjunto de datos. PC2 explica otro 22%.

Juntas, explican un 57% de la varianza total, lo que quiere decir que más de la mitad de la información original ya está representada en esas dos componentes.

  
Esta imagen muestra cómo cambia la agrupación de personas según su edad y horas trabajadas al aplicar k-means con distintos valores de **k** (número de grupos que tomamos). Con **k=2** no se distinguen bien las diferencias de horas trabajadas y los grupos de edad. A medida que se aumenta **k a 4 y luego a 10**, los grupos capturan más detalles, separando mejor a las personas según combinaciones específicas de edad y horas trabajadas.  
  
  
  
  
  
Este gráfico, de esta distribución, nos muestra como los ocupados son aquellos que tienen los mayores años de educación y los que a mayor edad, todavía con sus años de educación, se mantienen ocupados. Se puede ver en el grafico como los mayores valores de años de educación forman parte de los ocupados.

Mientras que los desocupados están más concentrados en bajos años de educación y en bajas edades aunque esto no quiere decir que no haya desocupados con muchos años de educación o con edades más grandes.