UNIVERSIDAD CATOLICA BOLIVIANA “SAN PABLO”

MAESTRIA EN CIENCIA DE DATOS, TERCERA VERSION



Materia: ANALISIS ESTADÍSTICO II

Practica No.3

Maestrante: Ramón Wilder Serdán Cárdenas

Julio 2022

La Paz – Bolivia

Contenido

[1.- Realizar un ejercicio de Árboles de Decisión 3](#_Toc108765940)

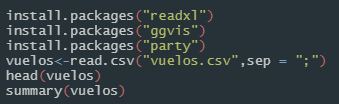
[2.- Realizar un ejercicio de Análisis de Componentes Principales. 7](#_Toc108765941)

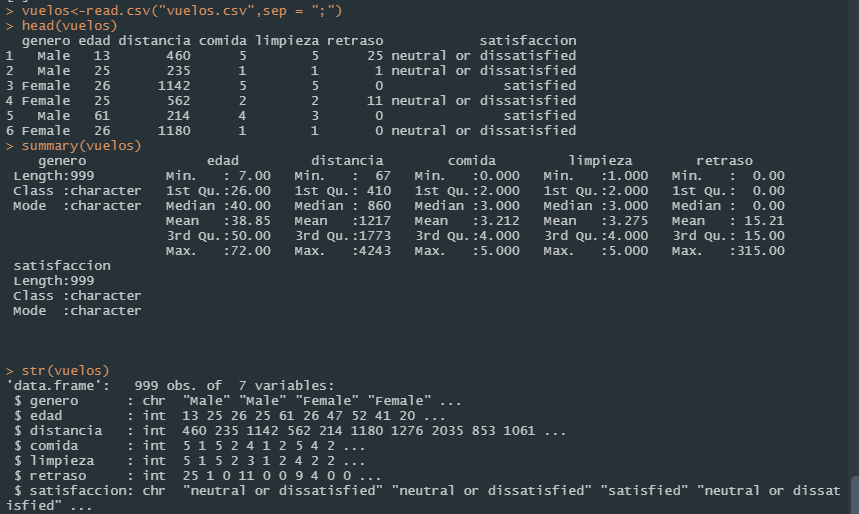
## 1.- Realizar un ejercicio de Árboles de Decisión

Se cuenta con información de una encuesta de satisfacción de mil usuarios de tren, los cuales están clasificados por Genero, Edad, Distancia del recorrido en kilómetros, la percepción de la Comida en el tren, la Limpieza y el Retraso en minutos de cada viaje.

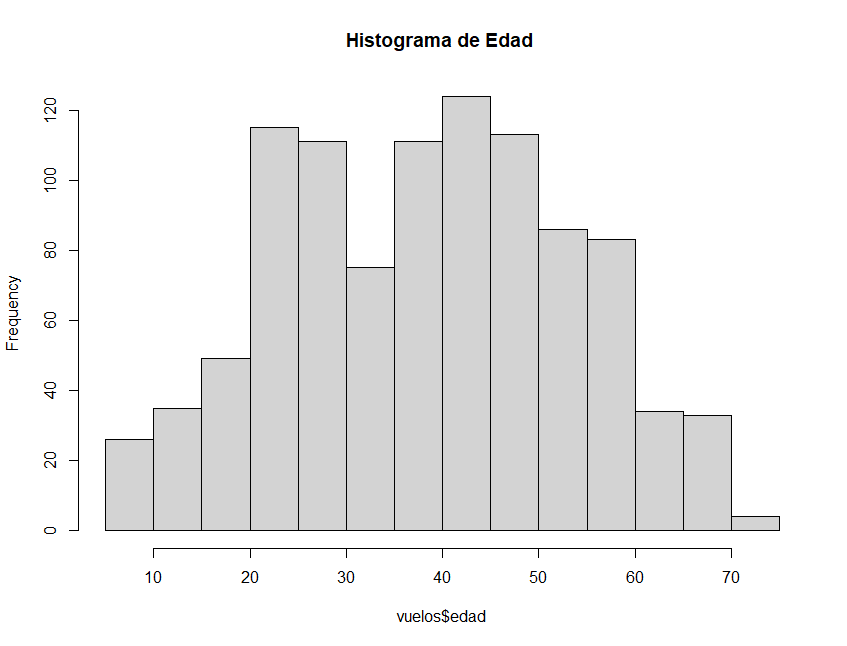
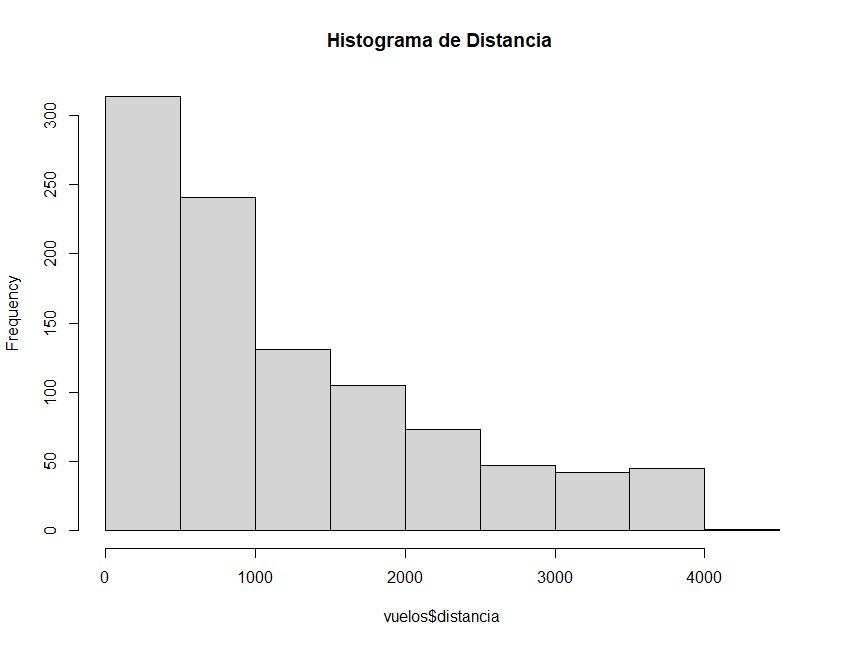
Se requiere un análisis de árboles de decisión para conocer a qué grupo de edad perteneces las diferentes variables de la encuesta.

Inicialmente importamos la información con ayuda de la librería “readxl” y mostramos algunos indicadores:



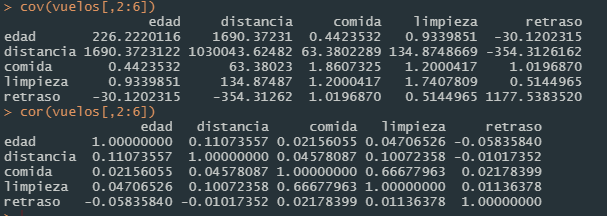


Visualizamos algunos histogramas:

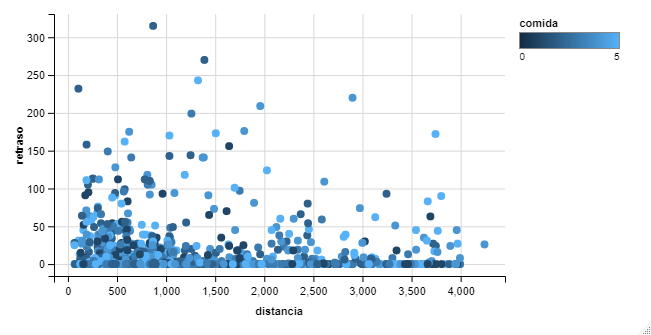
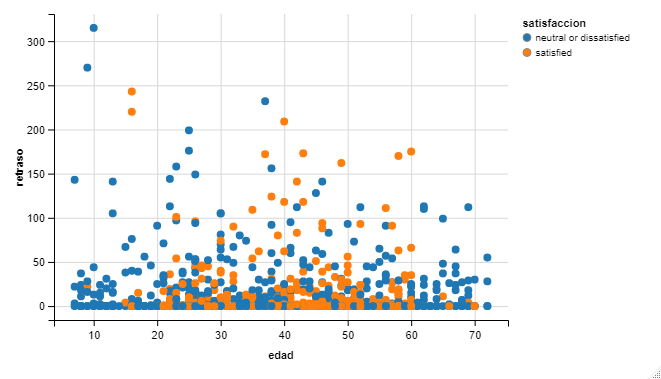


De las anteriores variables cuantitativas, sólo la edad de los pasajeros muestra una distribución normal.

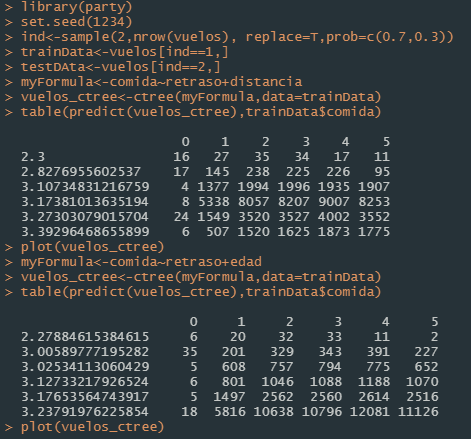
Asimismo, detallamos las matrices de covarianza y correlaciones:

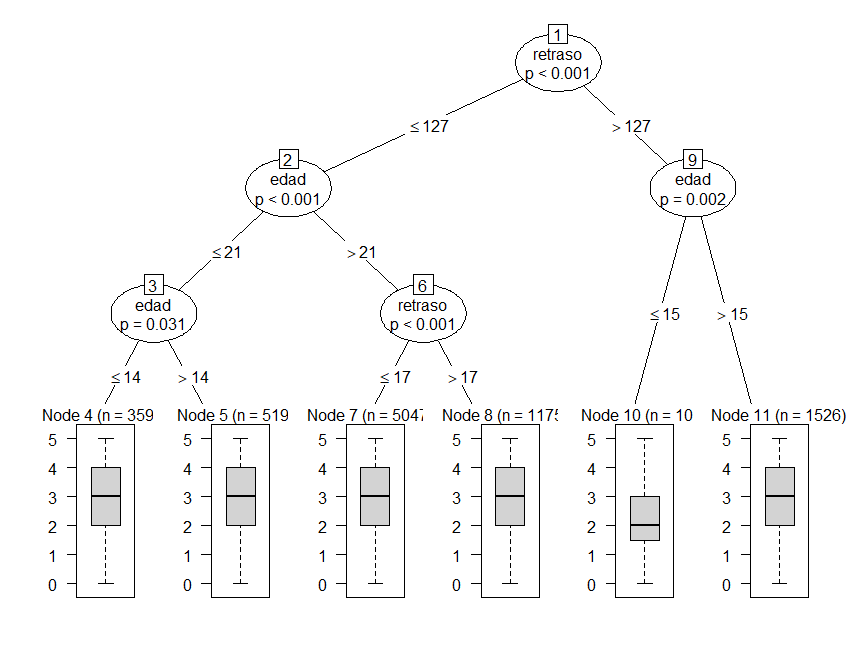


Gráficamente, evidenciamos la relación de algunas variables cuantitativas.



Aplicamos los comandos para la visualización del árbol de decisión con un dataset de entrenamiento de 70% de las observaciones y 30% de dataset de test.





El retraso en minutos es el primer factor para el análisis, seguido de la edad. Se concluye que el nivel de satisfacción de la comida es mayor en aquellos viajes que presentan un menor retraso. Sólo en el caso de los mayores de 15 años y un retraso importante de viajes, se presenta un nivel de satisfacción aceptable.

## 2.- Realizar un ejercicio de Análisis de Componentes Principales.

Se analizarán mediante componentes principales la información de 150 localidades de Bolivia en un relevamiento de información en el año 2011. Dicho año muestra información antes de la implementación plena de políticas económicas desentralizadas y participación activa del gobierno central. Las variables disponibles son:

Tasa de **desempleo,** es como la proporción de las personas en edad laboral que no tienen empleo, dispuestas a trabajar, y que han realizado acciones específicas para encontrar empleo.

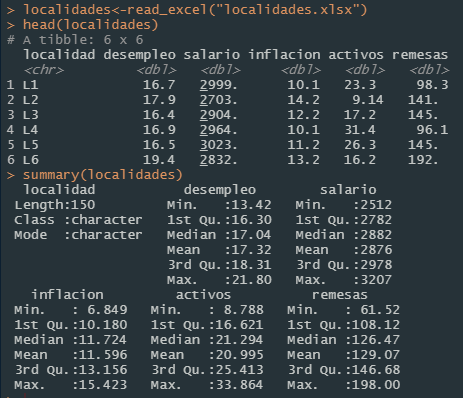
**Salario**, monto en bolivianos ganados en el tercer trimestre de dicha gestión.

**Inflación**, tasa de crecimiento de los precios medidos a través del Índice de Precios al Consumidor publicado por el INE.

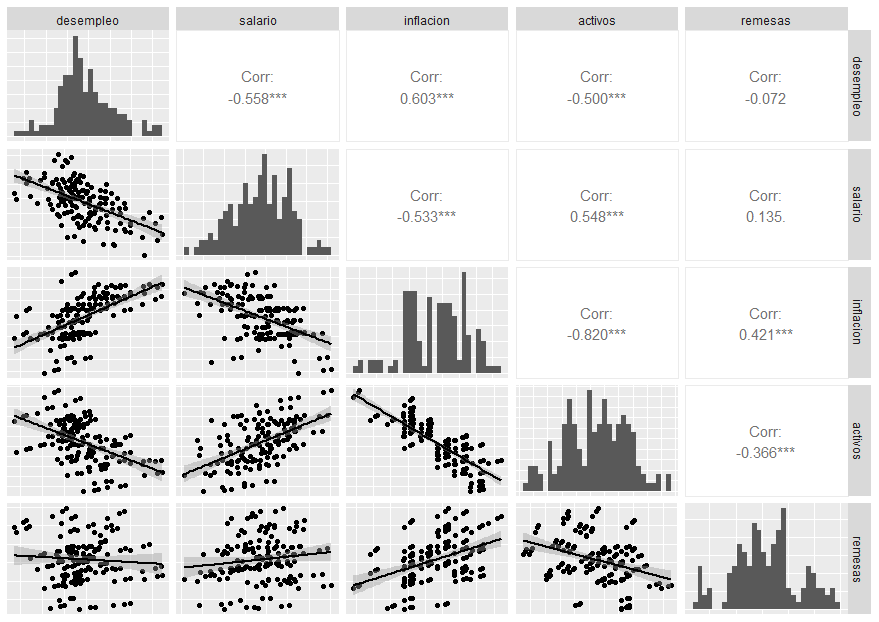
**Activos**, monto en millones de bolivianos imputados al territorio de la localidad en estudio como activos financieros.

**Remesas**, monto en miles de bolivianos recibidos de no residentes en el extranjero y destinados a gastos corrientes de los beneficiarios.

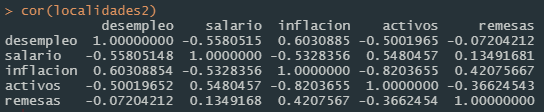
Leemos los datos para describir algunos indicadores de tendencia central.

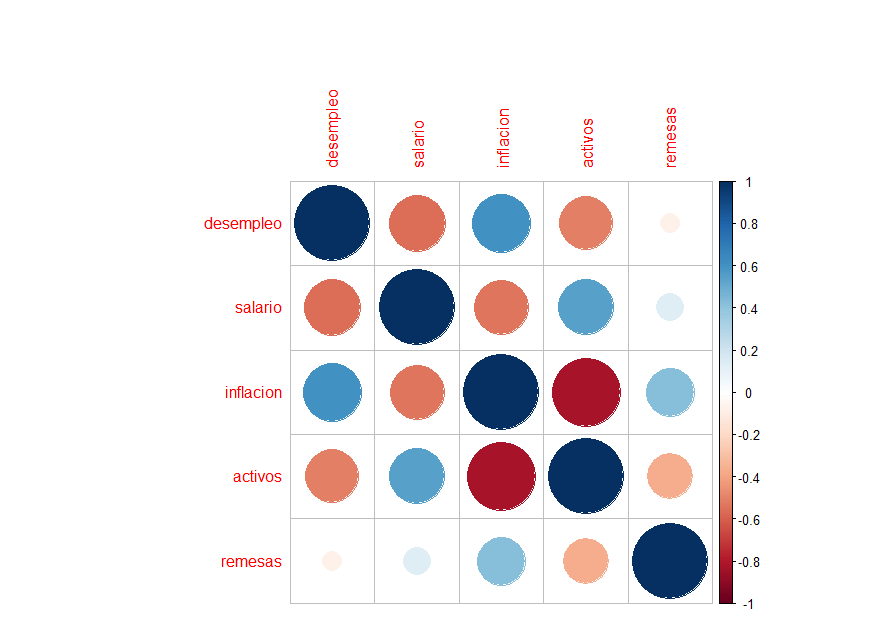


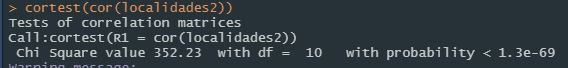
Asimismo, mostramos los histogramas de todas las variables cuantitativas evidenciando su comportamiento normal. Existe una alta correlación entre activos e inflación, seguido de inflación con desempleo.



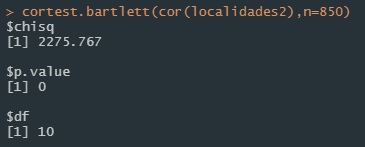
La matriz de correlaciones es la siguiente:

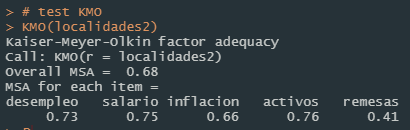




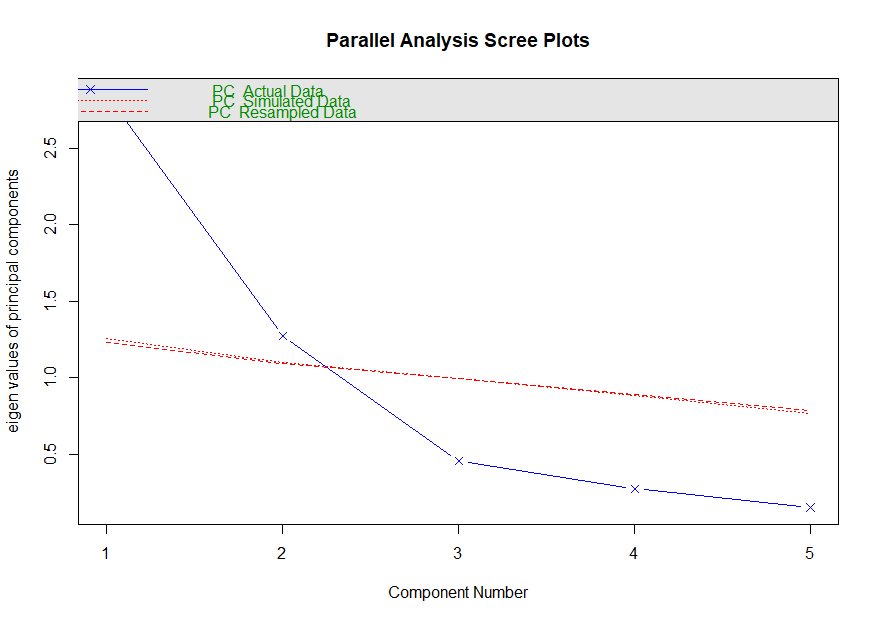
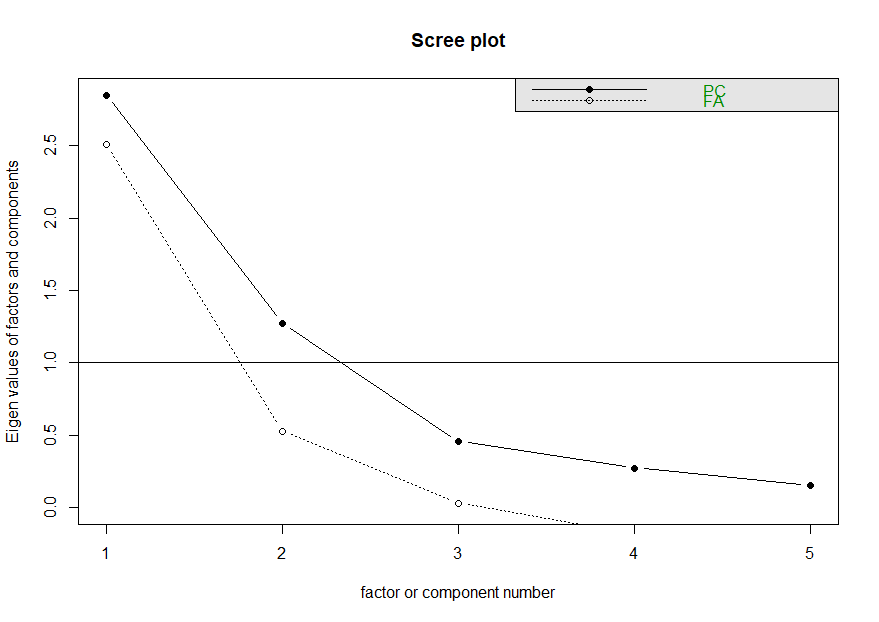


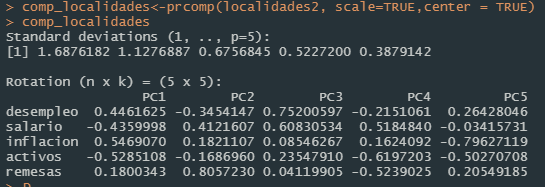
Se acepta la hipótesis alternativa de que las correlaciones son distintas.

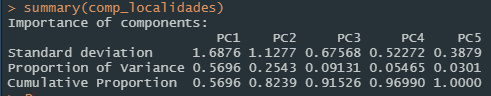




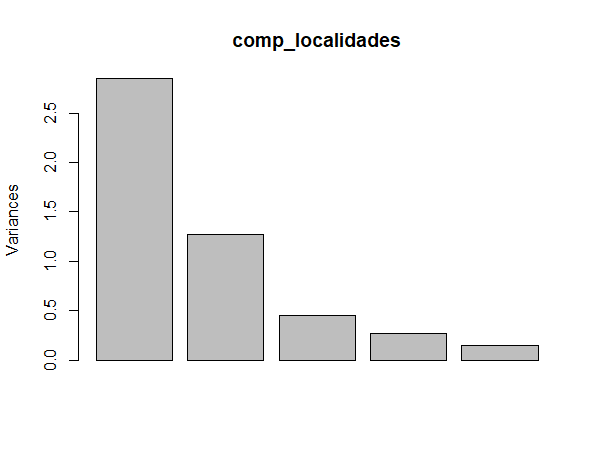
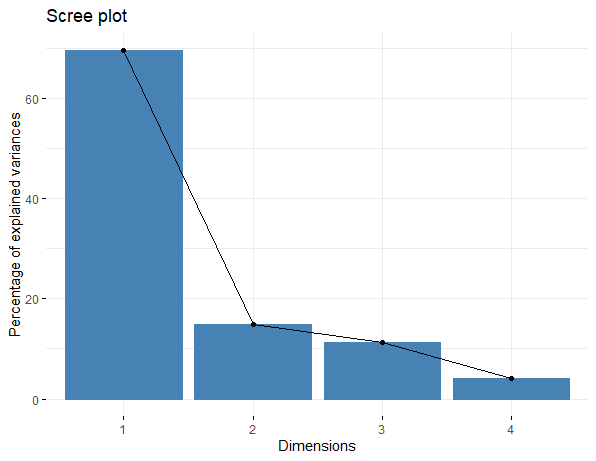
Mediante el test KMO justificamos elaborar análisis de componentes principales.



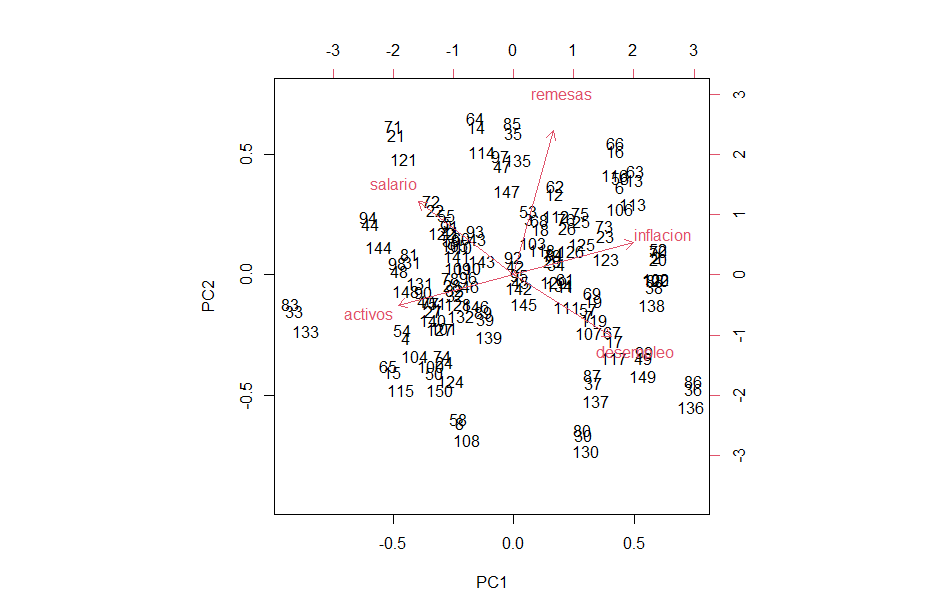
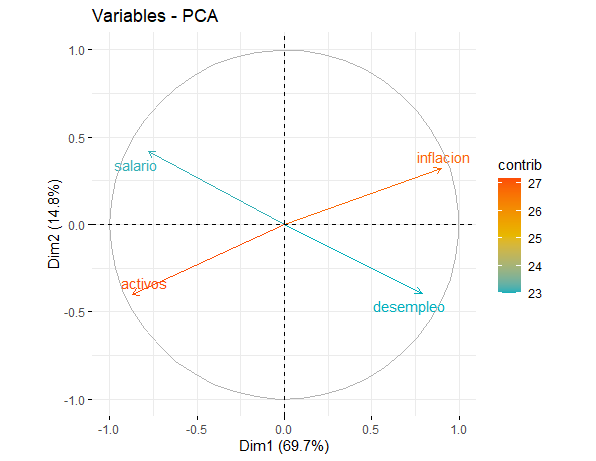
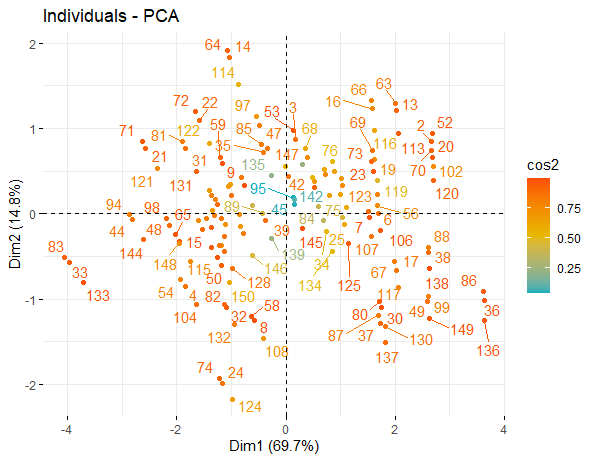


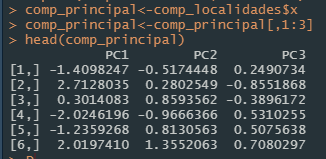


Mediante una representación gráfica obtenemos:



En un análisis gráfico preliminar podemos evidenciar una primera aproximación de los grupos principales.





Mediante el biplot visualizamos los resultados del Análisis de Componentes Principales ya que combina las puntuaciones de los componentes principales y los vectores de carga. En el mismo, las observaciones están etiquetadas por el código de localidad. La posición en el gráfico representa las puntuaciones de los dos primeros componentes principales. Las variables originales se muestran como vectores (flechas). El peso de la variable Desempleo en el primer componente es 0,5 y su peso en el segundo componente principal -0,99.