Proyecto Final

Heidy Alejandra Orjuela Ramírez

Ingeniería Estadística - Escuela Colombiana de Ingeniería Julio Garavito Nombre asignatura

Análisis Exploratorio de Datos Automatizado en R – library(amtv)

1 Introducción

El análisis exploratorio de datos (EDA, por sus siglas en inglés) es una etapa fundamental en cualquier proyecto de ciencia de datos, ya que permite obtener una comprensión inicial de la estructura, patrones y anomalías del conjunto de datos antes de aplicar modelos estadísticos o algoritmos de aprendizaje automático. Esta fase fue introducida formalmente por John Tukey en la década de 1970, quien planteó que el análisis estadístico debía ser un proceso iterativo e interactivo centrado en descubrir lo que los datos pueden decir por sí mismos.

EDA incluye tareas como la identificación de variables numéricas y categóricas, el análisis de valores faltantes, la detección de outliers, el cálculo de estadísticas descriptivas (media, mediana, desviación estándar, etc.), y la visualización de distribuciones y relaciones entre variables mediante gráficos. Además, herramientas como el análisis de correlación permiten establecer dependencias lineales y no lineales, mientras que técnicas de reducción de dimensionalidad, como el Análisis de Componentes Principales (PCA), permiten simplificar datasets complejos sin perder demasiada información relevante.

En el contexto actual, con la disponibilidad de herramientas estadísticas avanzadas y lenguajes como R, es posible automatizar muchas de estas tareas. Paquetes como knitr permiten generar reportes dinámicos en formatos como HTML o PDF que documentan todo el proceso de análisis, facilitando tanto la reproducibilidad como la comunicación de los resultados.

El presente proyecto se enfoca en implementar un conjunto de funciones en R para realizar un análisis exploratorio automatizado. Estas funciones están diseñadas para ser reutilizables y modulares, permitiendo generar informes detallados y visualizaciones que facilitan la comprensión del conjunto de datos, siendo de gran utilidad tanto en investigación como en entornos profesionales.

2 Objetivos

El propósito de la práctica fue desarrollar un paquete en R y aplicar funciones que permitan automatizar el análisis exploratorio de datos y análisis multivariado, generando reportes detallados que ayuden a identificar patrones, distribuciones y relaciones en los datos sin necesidad de realizar cálculos manuales extensos.

3 Material Utilizado

1. Lenguaje de programación: R
2. Paquetes principales:

Paquete	¿Para qué se usa?
`dplyr`	Manipulación de datos (`select` , `mutate` , `arrange` , `filter` , etc.)
`tidyr`	Conversión entre formatos anchos/largos (implícito con `as.table`)
`stats`	Función `cor()` para calcular matrices de correlación
`stringr`	Construcción de cadenas HTML (si decides usar `str_*` , opcional)

3. Sistema operativo: Windows 10, 64-bit
4. Computador: Laptop con procesador Intel Core i5, 8GB RAM
5. Referencias bibliográficas:
    - [1] J. F. Kurose y K. W. Ross, *First Chapter*, Pearson, 2022.
    - [2] J. Postel, “Internet protocol,” STD 5, RFC Editor, 1981.

4 Metodología

Se desarrollaron funciones personalizadas en R para automatizar el análisis exploratorio de datos. Primero, se diseñó una función principal (inf_gen) encargada de cargar el conjunto de datos y generar un resumen inicial, identificando automáticamente las variables numéricas y categóricas. A continuación, se implementaron funciones auxiliares como extraer_continuas para filtrar únicamente las variables numéricas continuas y inf_num para calcular estadísticas descriptivas detalladas (media, mediana, desviación estándar, entre otras), las cuales se presentan en formato HTML. Para evaluar relaciones entre variables, se programaron las funciones cor_pearson y cor_spearman, que calculan coeficientes de correlación usando ambos métodos y generan tablas dinámicas coloreadas según la fuerza de la relación. Además, se desarrolló la función red_dim, que aplica técnicas de reducción de dimensionalidad, como PCA, para detectar redundancias o patrones en los datos, luego de esto esta la función de gra_codo que nos ayuda a visualizar cuantos clusterings vamos a usar y para finalizar esta la función k_means_pca que nos ayuda a realizar la gráfica de cluster, teniendo en cuanta la reducción de dimensionalidad, Finalmente, se integraron todas estas funciones en un flujo automatizado que genera reportes HTML estructurados, facilitando la interpretación visual y cuantitativa de los resultados del EDA.

5 Resultados

La ejecución de las funciones permitió identificar las características principales de los datos, mostrando distribuciones, valores faltantes y correlaciones significativas. La automatización facilitó la obtención rápida de un reporte completo sin errores manuales, ayudando a detectar relaciones clave entre variables y posibles anomalías. Las visualizaciones generadas mejoraron la comprensión del dataset y la toma de decisiones para análisis posteriores.

1 FUNCIÓN

```
library(amtv)
library(nycflights13)

inf_gen(flights)
```

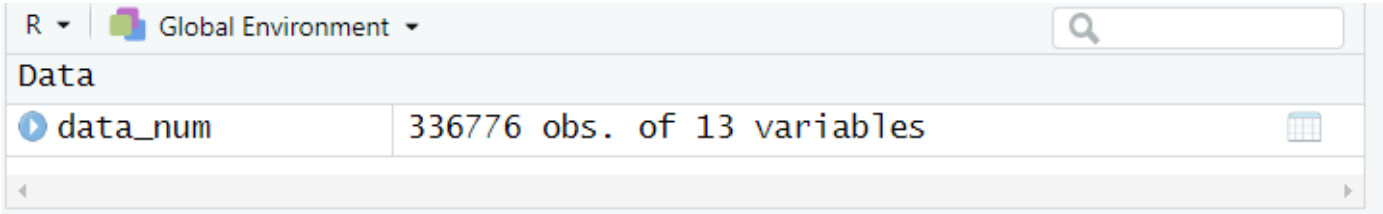
Informe EDA General

Tipos de variables

```
## $year
## [1] "integer"
##
## $month
## [1] "integer"
##
## $day
## [1] "integer"
##
## $dep_time
## [1] "integer"
##
## $sched_dep_time
## [1] "integer"
##
## $dep_delay
## [1] "numeric"
##
## $arr_time
## [1] "integer"
##
## $sched_arr_time
## [1] "integer"
##
## $arr_delay
## [1] "numeric"
##
```


2 FUNCIÓN

```
extraer_continuas(flights)
```



3 FUNCIÓN

```
inf_num(data_num)
```

Información General: Variables Numéricas

- Resumen Estadístico
- Valores Nulos por Variable
- Valores Atípicos
- Boxplots por Variable

Resumen Estadístico

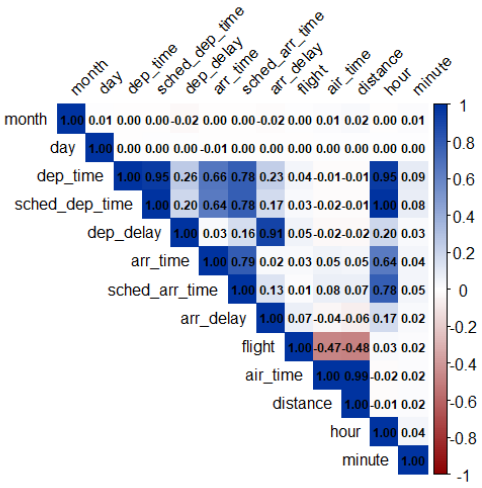
```
##      month             day           dep_time
##  Min.   : 1.000   Min.   : 1.00   Min.   :   1
##  1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.: 907
##  Median : 7.000   Median :16.00   Median :1401
##  Mean   : 6.549   Mean   :15.71   Mean   :1349
##  3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:1744
##  Max.   :12.000   Max.   :31.00   Max.   :2400
##                                   NA's   :8255
##  sched_dep_time   dep_delay          arr_time
##  Min.   : 106   Min.   : -43.00   Min.   :   1
##  1st Qu.: 906   1st Qu.:  -5.00   1st Qu.:1104
##  Median :1359   Median :  -2.00   Median :1535
##  Mean   :1344   Mean   :  12.64   Mean   :1502
##  3rd Qu.:1729   3rd Qu.:  11.00   3rd Qu.:1940
##  Max.   :2359   Max.   :1301.00   Max.   :2400
##                 NA's   :8255      NA's   :8713
##  sched_arr_time   arr_delay           flight
##  Min.   :   1   Min.   : -86.000   Min.   :   1
##  1st Qu.:1124   1st Qu.: -17.000   1st Qu.: 553
##  Median :1556   Median :  -5.000   Median :1496
##  Mean   :1536   Mean   :   6.895   Mean   :1972
##  3rd Qu.:1945   3rd Qu.:  14.000   3rd Qu.:3465
```

4 FUNCIÓN

```
cor_pearson(data_num)
```

Plot Zoom

Matriz de Correlación de Pearson

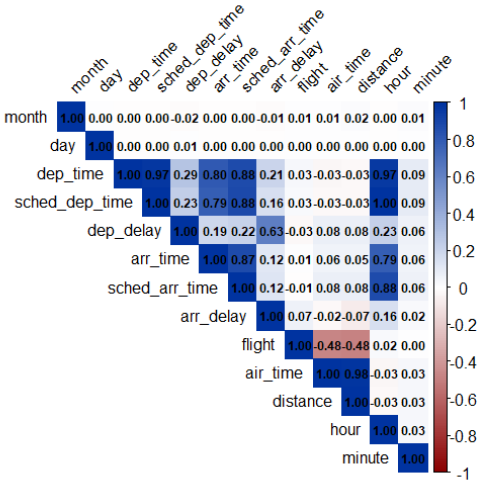


5 FUNCIÓN

```
cor_spearman(data_num)
cor_spearman(data_num, vars = c("air_time", "dep_delay
```

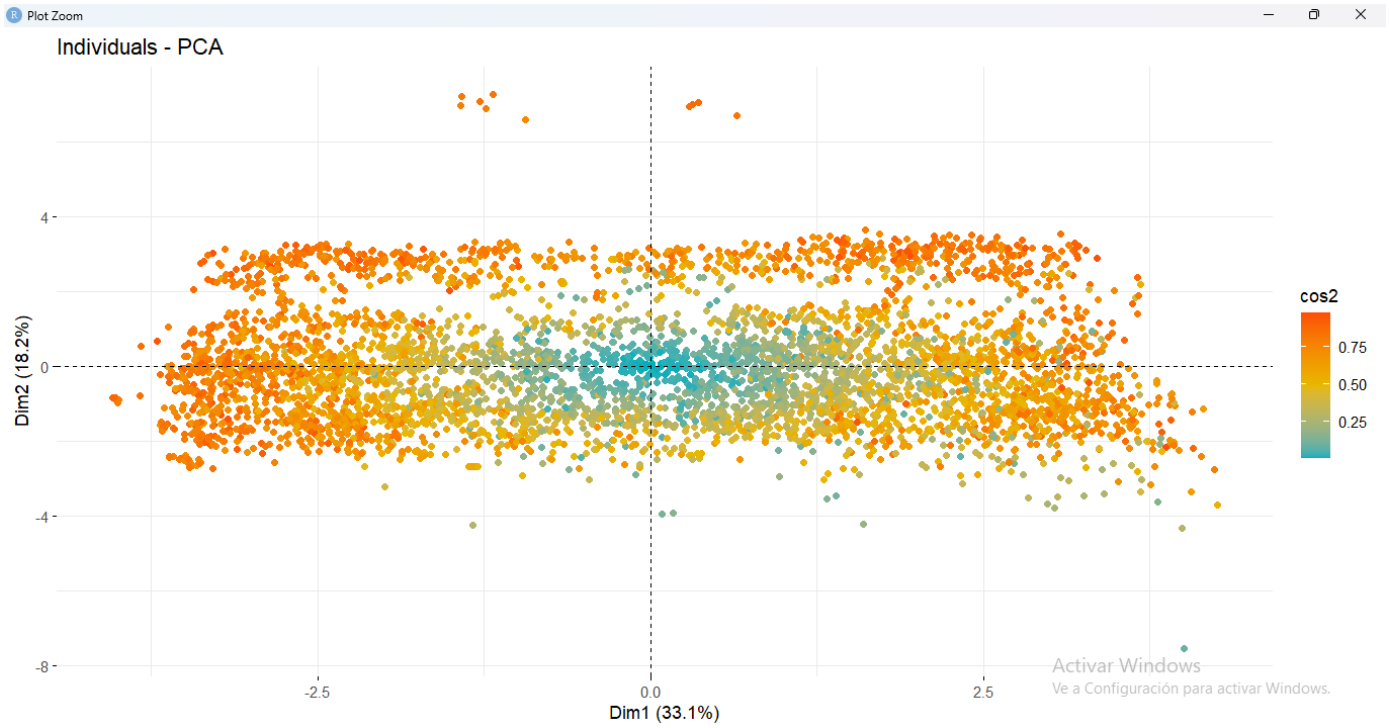
Plot Zoom

Matriz de Correlación de Spearman



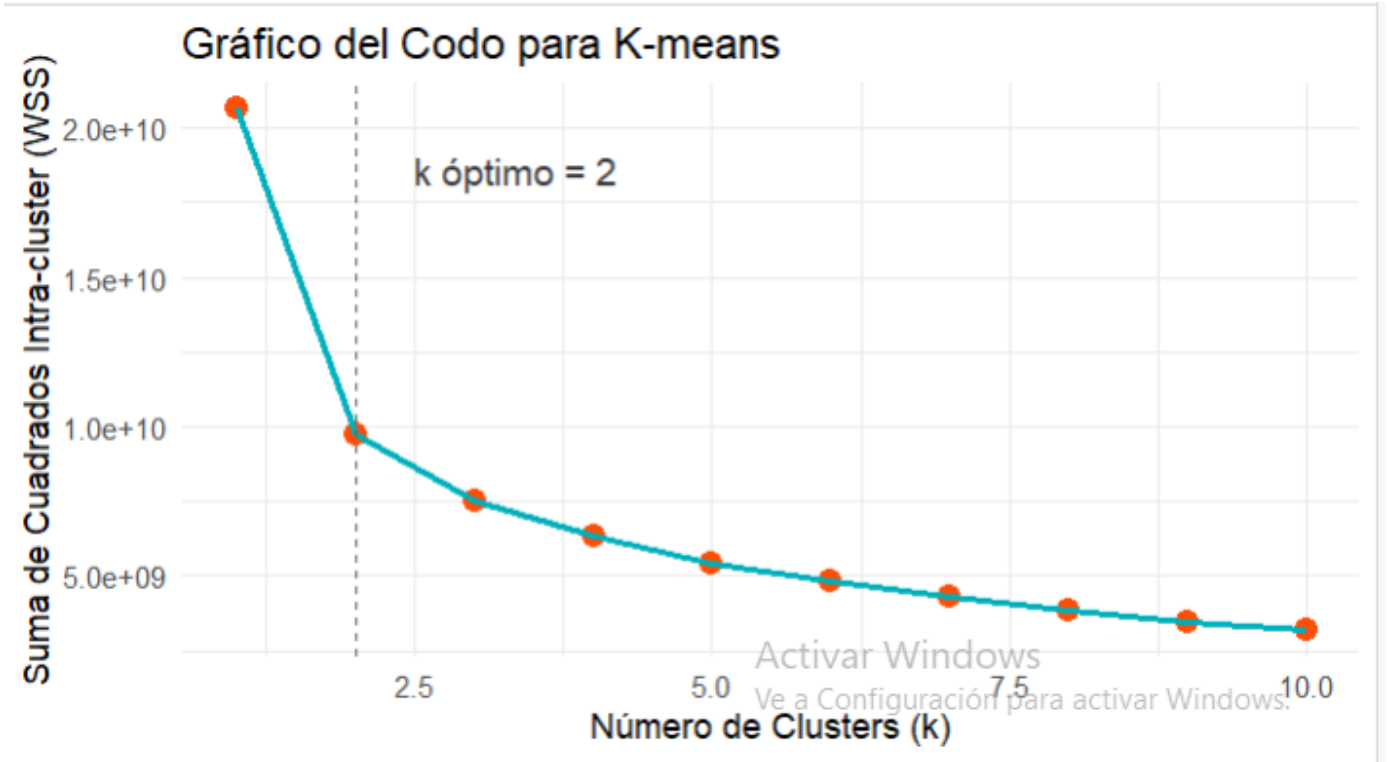
6 FUNCIÓN

```
red_dim(data_num)
```



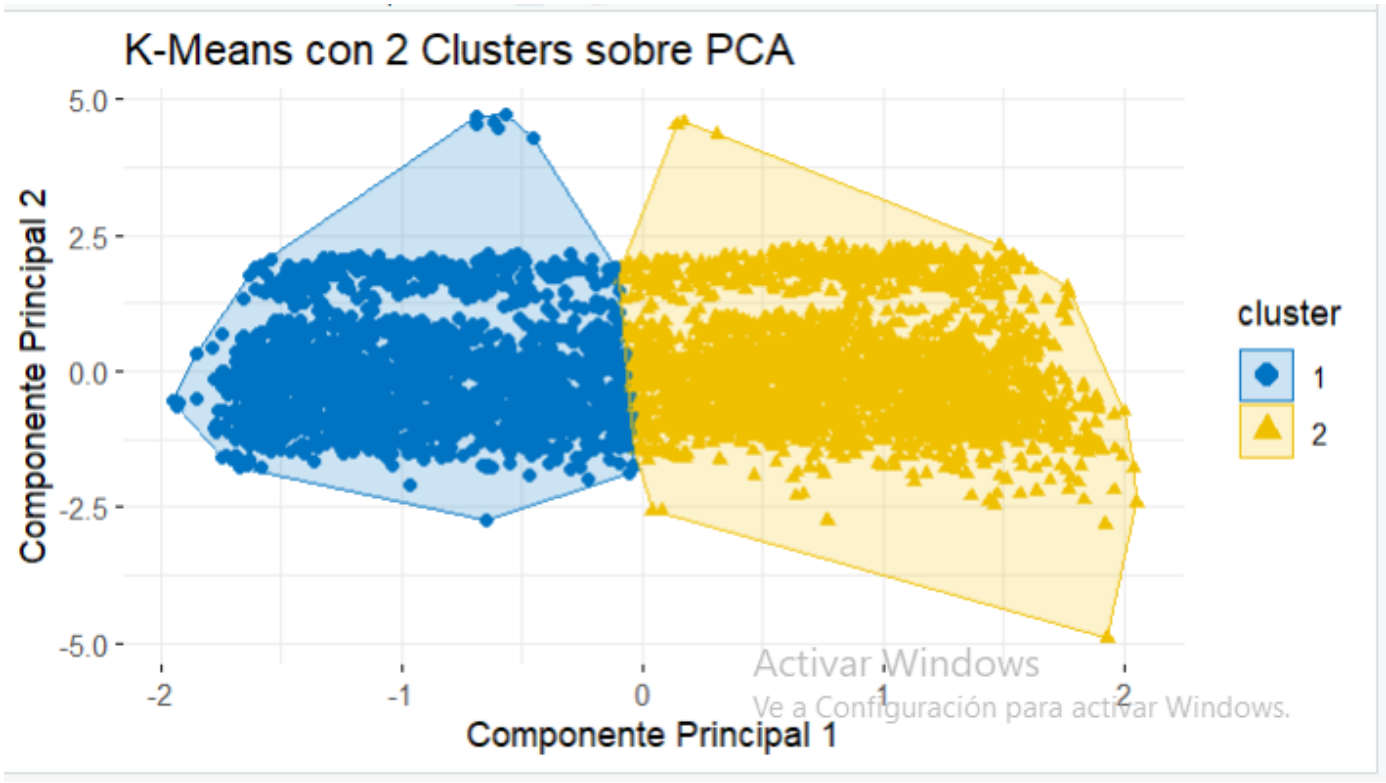
7 FUNCIÓN

```
gra_codo(data_muestra)
```



8 FUNCIÓN

```
k_means_pca(data_muestra)
k_means_pca(data_muestra, k=3)
```



6 Conclusiones

El proyecto demostró que es posible agilizar significativamente el análisis exploratorio de datos mediante el diseño e implementación de funciones automatizadas en R. Al encapsular tareas comunes del EDA —como la detección de variables numéricas, el cálculo de estadísticas descriptivas, la estimación de correlaciones y la reducción de dimensionalidad— en funciones reutilizables, se logró disminuir el tiempo invertido en procesos manuales, minimizar errores humanos y estandarizar los resultados obtenidos. Esto no solo optimiza el flujo de trabajo del analista de datos, sino que también facilita la escalabilidad del análisis cuando se trabaja con múltiples conjuntos de datos o en entornos colaborativos.

Además, la incorporación de la generación automática de informes en formato HTML representa un valor añadido importante, ya que permite presentar los hallazgos de manera clara, ordenada y visualmente accesible para distintos públicos, incluidos tomadores de decisiones no técnicos. Estos informes son, además, completamente reproducibles, lo que garantiza la trazabilidad de los resultados y favorece las buenas prácticas en ciencia de datos. En conjunto, el desarrollo de este conjunto de herramientas refuerza la idea de que el uso de programación funcional y automatización en R puede mejorar sustancialmente la eficiencia, precisión y comunicación en proyectos de análisis exploratorio de datos.

References

- Wickham, H. (2019). *Advanced R* (2nd ed.). Chapman and Hall/CRC.
- Grolemund, G., & Wickham, H. (2017). *R for data science*. O'Reilly Media.
- Peng, R. D. (2016). *R programming for data science*. Leanpub.
- Kuhn, M., & Johnson, K. (2019). *Feature engineering and selection: A practical approach for predictive models*. CRC Press.
- Marwick, B., Boettiger, C., & Mullen, L. (2018). Packaging data analytical work reproducibly using R (and friends). *The American Statistician, 72*(1), 80–88. https://doi.org/10.1080/00031305.2017.1375986
- Xie, Y. (2021). *Dynamic documents with R and knitr* (2nd ed.). Chapman and Hall/CRC.
- Müller, K., & Wickham, H. (2021). *R packages*. O'Reilly Media.
- Biecek, P. (2018). *Explanatory model analysis*. Chapman and Hall/CRC.