**INFORME ESTADISTICA DESCRIPTIVA**

# RESUMEN

El análisis exploratorio de datos es fundamental en cualquier estudio estadístico, por encima de análisis de aprendizaje de modelos estadísticos que sean capaces de predecir valores futuros, ya que si este paso no se hace con cautela no se tendrán datos de calidad y, por tanto, el modelo aprenderá de forma ineficiente.

En esta práctica tomaremos un primer contacto con las técnicas estadísticas que se utilizan para ordenar, analizar y representar un conjunto de datos, con el fin de describir apropiadamente sus características. El primer paso en el análisis de datos, una vez introducidos los mismos, es realizar un análisis descriptivo o análisis exploratorio de datos. Los principales objetivos de un análisis descriptivo o análisis exploratorio de datos son la descripción y la síntesis de los datos. Para ello, los datos se organizan en tablas, se calculan medidas que describen sus características más importantes y se realizan representaciones gráficas.

# PACKAGE DE DATOS

**Package ‘insuranceData’.-** Conjuntos de datos de seguros, que a menudo se usan en la severidad de las reclamaciones y el modelado de frecuencia de reclamos. Ayuda a probar nuevos modelos de regresión en esos problemas, como GLM, GLMM, HGLM, modelos mixtos no lineales, etc.

# DESCRIPCIÓN DEL EJERCICIO

Se toma como muestra el conjunto de datos de:

**data(dataCar) .-** Este conjunto de datos se basa en pólizas de seguro de un año tomadas en 2004 o 2005. Hay 67856 pólizas, de las cuales 4624 se aseguraron.

**Formato de datos:**

Un marco de datos con 67856 observaciones sobre las siguientes 11 variables.

**veh\_value.-** valor del vehículo, en $ 10,000

**exposure** .- 0-1

**clm**.- ocurrencia del reclamo (0 = no, 1 = sí)

**numclaims**.- número de reclamos

**claimcst0**.- monto de la reclamación (0 si no hay reclamo)

**veh\_body**.- cuerpo del vehículo, codificado como BUS CONVT COUPE HBACK HDTOP MCARA MIBUS PANVN RDSTR SEDAN STNWG TRUCK UTE

**veh\_age.-** 1 (más joven), 2, 3, 4

**gender**.- un factor con niveles F M

**área**.- un factor con niveles A B C D E F

**agecat**.- 1 (más joven), 2, 3, 4, 5, 6

**X\_OBSTAT\_.-**un factor con niveles 01101 0 0 0

1. **ANÁLISIS**

**Librerías utilizadas:**

install.packages('insuranceData')

library(insuranceData)

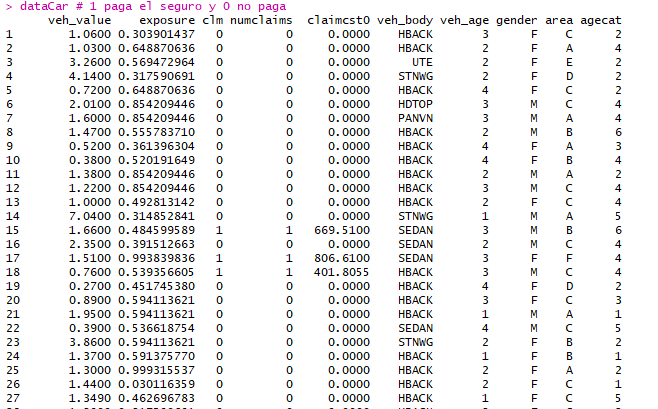
library(PASWR)

install.packages("ggplot2")

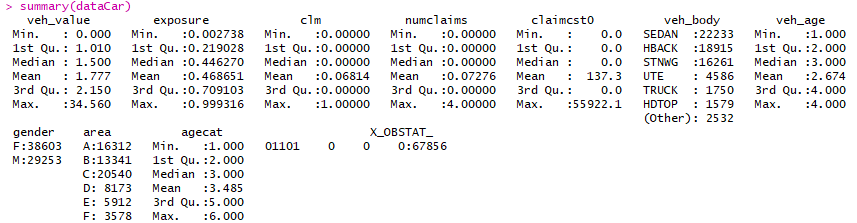
library(ggplot2)

**Datos:**

data(dataCar)

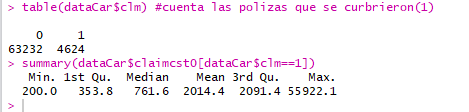


Realizamos un resumen estadístico de los datos, donde se observa que existen datos no reales ya que se encuentran datos atípicos, el costo de una póliza de seguro como mínima no puede ser cero y el máximo es $55922,1.



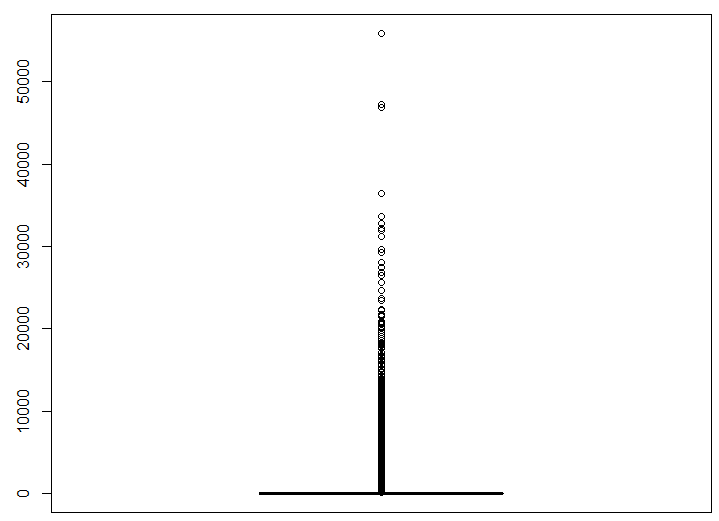
Distinguimos cuantos datos atípicos existen:

Viene a ser el valor promedio de los valores que pagaron el seguro sin los atípicos.



Realizamos la gráfica para observar los datos:





Grafica con el total de los datos

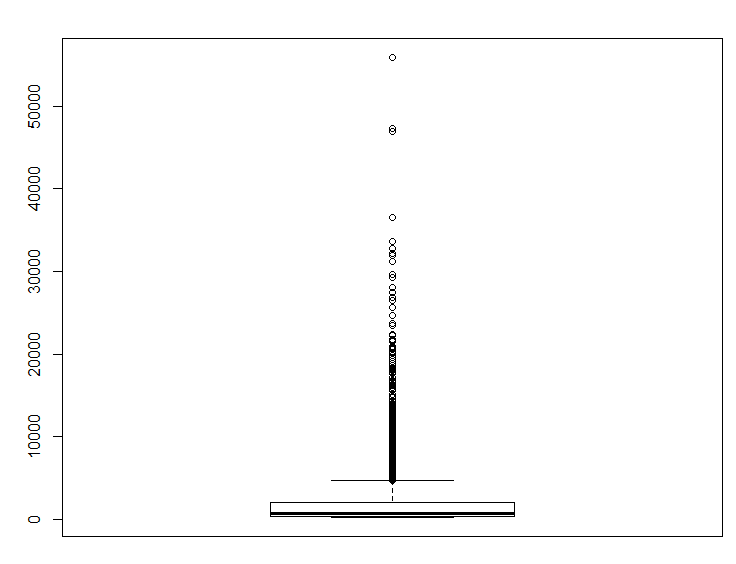


Gráfico solo con datos que pagaron el seguro.

Número de datos atípicos que se extraen.





