

# Trabajo final Estadística Avanzada

Carlos Alberto Murillo M

Luz Stella Florez  
Cindy Guerra

Diana Carolina Benjumea

noviembre 28, 2020



# Contents

<b>Abstract</b>	<b>1</b>
<b>Objetivos y Lineamientos</b>	<b>3</b>
<b>Capítulo 1. Lectura de variables de empresa</b>	<b>5</b>
Selección de las fuentes de información . . . . .	5
Proceso de carga de los datos . . . . .	6
<b>Capítulo 2. Lectura y consolidación de variables económicas</b>	<b>11</b>
<b>Capítulo 3. Consolidación de la base</b>	<b>15</b>
Estandarización de variables . . . . .	25
<b>Capítulo 5. Correlaciones</b>	<b>27</b>
Calcular el coeficiente de correlación . . . . .	28
Visualizar matriz de correlación . . . . .	30
<b>Capítulo 6. Aplicación de modelo</b>	<b>33</b>
6.1. Modelo regresión Lineal - con todas las variables de compra y venta de oro .	33
6.1. Modelo regresión Lineal . . . . .	35
6.2. Modelo Regresión lineal sin efectos aleatorios . . . . .	36
6.3. Modelo lineal con intercepto aleatorio . . . . .	38
6.4. Modelo Lineal con intercepto aleatorio a nivel de NIT . . . . .	39
<b>Capítulo 7. Estimación de esfuerzo</b>	<b>43</b>
<b>Capítulo 8. Conclusiones</b>	<b>45</b>
<b>REFERENCIAS</b>	<b>47</b>



# Abstract

Este documento analiza los impactos de las variables macroeconómicas en los costos y gastos de empresas del sector “Extracción de oro y otros metales preciosos”. Para el análisis, se tomaron los datos de los años: 2017, 2018 y 2019. El código de este trabajo se encuentra almacenado en el repositorio de Github: [https://github.com/cabymetal/TrabajoFinal\\_estadisticos\\_avanzados](https://github.com/cabymetal/TrabajoFinal_estadisticos_avanzados).



# Objetivos y Lineamientos

Caracterizar las relaciones entre algunos indicadores macroeconómicos y los costos y gastos de ventas de las empresas colombianas vigiladas por la SuperSociedades.

Lineamientos:

1. Con ayuda de un modelo lineal modele cree un modelo o varios modelos que permitan caracterizar la relación entre las variables PIB, Inflación, Desempleo, Tasa de Cambio, Balance Fiscal, Balance en Cuenta Corriente, Tasa de intervención, TRM y los costos y gastos de ventas.
2. Se debe escoger mínimo un tipo de empresas (Clasificación Industrial Internacional Uniforme) que tenga más de 20 empresas y tomar al menos los últimos tres años de información disponible.
3. Se debe evaluar el ajuste y la capacidad predictiva.
4. Se deben explicar todas las transformaciones de variables requeridas por el modelo.
5. Se deben explicar todos los pasos para la construcción de la base de datos: descarga de información, concatenación, etc.
6. Se debe incluir un análisis descriptivo.
7. Se debe incluir un análisis de la razonabilidad de las cifras.
8. Se debe redactar un reporte técnico documentando lo anterior. La sugerencia es utilizar un formato que permita la inclusión de gráficos basados en html o JavaScript (por ejemplo hmtl a partir de Rmarkdown). El código se debe subir a un repositorio Git y referenciarlo en el reporte. El reporte debe incluir una estimación del esfuerzo de las actividades de 1) consolidación de información, 3) transformación de variables y análisis descriptivo, 4) ajuste y validación de modelos y 5) redacción del reporte.
9. El trabajo se debe subir al canal del curso en Teams y se debe notificar por correo a la dirección judaospi@bancolombia.com.co.
10. La fecha de entrega es el viernes 30 de octubre y el trabajo se puede presentar en equipos de máximo cinco estudiantes.

Para acceder a los datos de costos y gastos de ventas: • Entrar a <http://pie.supersociedades.gov.co> > MENÚ > Descarga Masiva de Información Descargar la información de los años 2016 a 2019





# Capítulo 1. Lectura de variables de empresa

## Selección de las fuentes de información

Para los datos básicos y financieros de las empresas, tomamos los siguientes archivos de la página de la Supersociedades:

- datosBasicosComplete.xlsx
- Plenas - Individuales2017.xlsx
- Plenas - Individuales2018.xlsx
- Plenas - Individuales2019.xlsx

Primera iteración:

Código CIIU seleccionado: G4711

Macrosector: Comercio

Descripción: Comercio al por menor en establecimientos no especializados con surtido compuesto principalmente por alimentos, bebidas o tabaco.

Esta clase incluye:

- Los establecimientos no especializados de comercio al por menor de productos cuyo surtido está compuesto principalmente de alimentos (víveres en general), bebidas o tabaco. No obstante, expenden otras mercancías para consumo de los hogares tales como vestuario, electrodomésticos, muebles, artículos de ferretería, cosméticos, entre otros. Suelen realizar este tipo de actividad los denominados supermercados, cooperativas de consumidores, comisariatos y otros establecimientos similares. También se incluyen las tiendas, los graneros, entre otros, que se encuentran en los pueblos o en barrios tradicionales.

Esta clase excluye:

- El expendio de comidas preparadas en restaurantes, cafeterías y por autoservicio.

Al realizar los cargues iniciales de información, nos dimos cuenta de que cruzaban muy pocas empresas, el conjunto de datos seleccionado no era suficiente, por lo que decidimos utilizar otro CIIU.

Segunda iteración:

Código CIIU seleccionado: B0722

Descripción: Extracción de oro y otros metales preciosos

Esta clase incluye:

- La extracción de oro, plata y otros metales del grupo del platino (osmio, iridio, rodio, rutenio y paladio).
- Las actividades realizadas para extraer el oro existente en los lechos de los ríos sin importar el sistema de extracción empleado (barequeo, motobombas, draguetas, dragas, elevadores, monitores u otros).
- La extracción de los metales preciosos se realiza a través de dos métodos: de veta o filón, que consiste en la extracción manual, mecanizada o semimecanizada de oro y de plata presentes en las rocas formando venas, vetas o filones.
- Las actividades o procesos físicos necesarios para separar el oro de la roca que lo contiene, conocidos como procesos de beneficio del mineral, de los cuales los más comunes son la trituration y la molienda (pulverización).
- Otros procesos tales como lavado (mazamorreo) hasta separar el oro y la plata de otros elementos o impurezas, siempre y cuando se realicen por cuenta del explotador y en sitios cercanos a la mina.
- El segundo método consiste en la extracción de oro o platino de aluviones (concentración de mineral en el lecho de los ríos), el cual se realiza por diferentes sistemas de extracción, tales como: barequeo (mazamorreo); pequeña minería, representada por grupos de trabajadores que utilizan motobombas, elevadores y draguetas; mediana minería, utilizando maquinaria como retroexcavadoras y buldózers, y la gran minería que realiza la extracción de metales preciosos por medio de dragas de cucharas.

Esta clase excluye:

- Los servicios de apoyo para la extracción de oro y metales preciosos. Se incluyen en la clase 0990, «Actividades de apoyo para otras actividades de explotación de minas y canteras».

## Proceso de carga de los datos

En este capítulo explicaremos el proceso de la carga de los datos de las empresas.

```
library(tidyverse)
library("readxl")
library("dplyr")
```

### 1. Cargamos los datos básicos de las empresas

```
#Revisamos como son nuestros datos para saber si tenemos que realizar algún
#ajuste a la carga
#file.show("./data/datosBasicosComplete.xlsx")

#Como el archivo no tiene forma de tabla al principio, debemos realizar la
#carga, ignorando las primeras filas del archivo.

#Cargar los archivos a un dataframe
pd_datos_basicos <- read_excel("./data/datosBasicosComplete.xlsx",
                               sheet = "Reporte", skip=8, col_types = c("text",
"text", "text", "text", "text", "text", "text", "text", "text", "text", "text",
"text", "text", "text", "text", "text", "date", "text", "date", "text", "date", "text",
```

```
"text"))

pd_datos_basicos %>%
  mutate(`Órgano Societario` = as.factor(`Órgano Societario`),
         `Etapla Situación` = as.factor(`Etapla Situación`)) -> pd_datos_basicos

head(pd_datos_basicos)

## # A tibble: 6 x 23
##   NIT   `Razón social` `Código CIIU` `Tipo Societari` `Objeto Social`
##   <chr> <chr>          <chr>          <chr>          <chr>
## 1 1001~ NOREÑA MANRI~ 0                PERSONA NATURAL <NA>
## 2 1001~ PEÑA RAMIREZ ~ H5229            PERSONA NATURAL <NA>
## 3 1002~ GONZALEZ SANC~ G4731            PERSONA NATURAL <NA>
## 4 1002~ RODRIGO JAVIE~ L6810            PERSONA NATURAL <NA>
## 5 1002~ BUITRAGO GONZ~ H4923            PERSONA NATURAL <NA>
## 6 1005~ KAREN JULIETH~ M7500            PERSONA NATURAL <NA>
## # ... with 18 more variables: `Dirección Notificación Judicial` <chr>, `Ciudad
## #   Notificación Judicial` <chr>, `Departamento Notificación Judicial` <chr>,
## #   `Teléfono Notificación Judicial` <chr>, `Dirección Domicilio` <chr>,
## #   `Ciudad Domicilio` <chr>, `Departamento Domicilio` <chr>, `Apartado
## #   Domicilio` <chr>, `E-Mail` <chr>, Web <chr>, Estado <chr>, `Fecha
## #   Estado` <dtm>, Situación <chr>, `Fecha Situación` <dtm>, `Etapla
## #   Situación` <fct>, `Fecha Etapla` <dtm>, `Nombre Representante Legal` <chr>,
## #   `Órgano Societario` <fct>
```

2. Filtramos los datos del CIIU seleccionado y de las empresas que se encuentren en situación activa.

```
library(dplyr)

pd_datos_basicos_flt <- pd_datos_basicos[,c("NIT", "Razón social", "Código CIIU",
      "Ciudad Domicilio", "Departamento Domicilio", "Estado", "Situación",
      "Órgano Societario", "Etapla Situación")]

names(pd_datos_basicos_flt) = c("NIT", "razon_social", "CIIU", "ciudad",
      "departamento", "estado", "situacion",
      "organo_societario", "etapla_situacion")

pd_datos_basicos_flt <- filter(pd_datos_basicos_flt, CIIU == "B0722" &
      situacion == "ACTIVA")

head(pd_datos_basicos_flt)

## # A tibble: 6 x 9
##   NIT   razon_social CIIU   ciudad departamento estado situacion organo_societar~
##   <chr> <chr>          <chr> <chr>   <chr>          <chr> <chr>      <fct>
## 1 8002~ GRUPO DE BU~ B0722 MEDEL~ ANTIOQUIA   INSPE~ ACTIVA   ACTIVIDAD ECONO~
## 2 8110~ MINERA CROE~ B0722 MEDEL~ ANTIOQUIA   INSPE~ ACTIVA   ACTIVIDAD ECONO~
## 3 8110~ NUEVA CALIF~ B0722 MEDEL~ ANTIOQUIA   INSPE~ ACTIVA   ACTIVIDAD ECONO~
## 4 8110~ COLOMBIA GO~ B0722 MEDEL~ ANTIOQUIA   INSPE~ ACTIVA   ACTIVIDAD ECONO~
## 5 8110~ NEGOCIOS MI~ B0722 MEDEL~ ANTIOQUIA   INSPE~ ACTIVA   ACTIVIDAD ECONO~
```

```
## 6 8300~ ECO ORO MIN~ B0722 BUCAR~ SANTANDER    INSPE~ ACTIVA    ACTIVIDAD ECONO~
## # ... with 1 more variable: etapa_situacion <fct>
```

### 3. Cargamos los datos financieros

```
pd_datos_fin_2017 <- read_excel("../data/Plenas - Individuales2017.xlsx",
                                sheet = "Estado de Resultado Integral" )

pd_datos_fin_2017 <- pd_datos_fin_2017[,c("Nit", "Periodo", "Costo de ventas",
"Costos de distribución", "Gastos de administración", "Otros gastos, por función",
"Costos financieros", "Gasto (ingreso) por impuestos, operaciones continuadas",
"Ingresos de actividades ordinarias", "Otros ingresos", "Ingresos financieros")]

names (pd_datos_fin_2017) = c("NIT", "Periodo", "costo_ventas",
    "costo_distribucion", "gastos_administracion", "otros_gastos",
    "costos_financieros", "gasto_impuestos_operaciones",
    "ingresos_actividades_ordinarias", "otros_ingresos", "ingresos_financieros")

datos_completos_fin <- merge (pd_datos_basicos_flt, pd_datos_fin_2017,
                              by.x="NIT", by.y="NIT")
```

Para efectos del ejercicio, no tomaremos el archivo de 2017, ya que el archivo 2018 tiene los datos de 2017 con la nueva norma.

```
pd_datos_fin_2018 <- read_excel("../data/Plenas - Individuales2018.xlsx",
                                sheet = "ERI" )

pd_datos_fin_2018 <- pd_datos_fin_2018[,c("Nit", "Periodo", "Costo de ventas",
"Gastos de ventas", "Gastos de administración", "Otros gastos",
"Costos financieros", "Ingreso (gasto) por impuestos",
"Ingresos de actividades ordinarias", "Otros ingresos", "Ingresos financieros")]

names (pd_datos_fin_2018) = c("NIT", "Periodo", "costo_ventas", "gastos_ventas",
    "gastos_administracion", "otros_gastos", "costos_financieros", "gasto_impuestos",
    "ingresos_actividades_ordinarias", "otros_ingresos", "ingresos_financieros" )

datos_completos_2018 <- merge (pd_datos_basicos_flt, pd_datos_fin_2018,
                              by.x="NIT", by.y="NIT")

#Le damos formato a los periodos

datos_completos_2018$Periodo[datos_completos_2018$Periodo == "Periodo Anterior"] <-
    "2017"
datos_completos_2018$Periodo[datos_completos_2018$Periodo == "Periodo Actual"] <-
    "2018"

pd_datos_fin_2019 <- read_excel("../data/Plenas - Individuales2019.xlsx",
                                sheet = "ERI" )

pd_datos_fin_2019 <- pd_datos_fin_2019[,c("Nit", "Periodo", "Costo de ventas",
"Gastos de administración", "Otros gastos", "Costos financieros",
```

```

    "Ingreso (gasto) por impuestos", "Ingresos de actividades ordinarias",
    "Otros ingresos", "Ingresos financieros"]

names(pd_datos_fin_2019) = c("NIT", "Periodo", "costo_ventas",
    "gastos_administracion", "otros_gastos", "costos_financieros",
    "gasto_impuestos", "ingresos_actividades_ordinarias", "otros_ingresos",
    "ingresos_financieros" )

datos_completos <- merge (pd_datos_basicos_flt, pd_datos_fin_2019,
    by.x="NIT", by.y="NIT")

datos_completos$Periodo[datos_completos$Periodo == "Periodo Actual"] <- "2019"

datos_completos <- filter(datos_completos, Periodo == "2019")

```

Se realizará el análisis con los periodos: 2017, 2018, 2019

```

#Eliminamos variable diferente a 2019
datos_completos_2018 <- select(datos_completos_2018, -gastos_ventas)

datos_completos = rbind(datos_completos, datos_completos_2018)
datos_completos %>% mutate( razon_social = as.factor(razon_social),
    CIIU = as.factor(CIIU), ciudad = as.factor(ciudad),
    departamento = as.factor(departamento),
    estado =as.factor(estado), Periodo = as.factor(Periodo),
    situacion = as.factor(situacion)) -> datos_completos

```

Utilizaremos las siguientes variables:

```
as.data.frame(colnames(datos_completos))
```

```

##          colnames(datos_completos)
## 1                      NIT
## 2          razon_social
## 3                      CIIU
## 4                  ciudad
## 5          departamento
## 6                  estado
## 7              situacion
## 8      organo_societario
## 9          etapa_situacion
## 10                 Periodo
## 11             costo_ventas
## 12      gastos_administracion
## 13             otros_gastos
## 14      costos_financieros
## 15             gasto_impuestos
## 16 ingresos_actividades_ordinarias
## 17             otros_ingresos
## 18      ingresos_financieros

```



## Capítulo 2. Lectura y consolidación de variables económicas

A continuación se presenta el proceso que se ejecutó para generar un dataframe con las variables: PIB, Inflación, Desempleo, Balance Fiscal, Balance en Cuenta Corriente, Tasa de intervención y TRM.

**PIB:** Es un indicador económico que refleja el valor monetario de todos los bienes y servicios finales producidos por un país o región en un determinado periodo de tiempo, normalmente un año. Se utiliza para medir la riqueza que genera un país. Para cada año se tomó el promedio de los datos diarios. Datos tomados de aquí.

```
library(dplyr)
#Los datos son tomados de https://datosmacro.expansion.com/pib/colombia
# vectores
anyo <- c("2016", "2017", "2018", "2019")
PIB_M.E. <- c(289.239, 280.249, 275.999, 255.416)
Var.PIB <- c(3.3, 2.5, 1.4, 2.1)
#Crear dataframe de vectores
PIB <- data.frame(anyo, PIB_M.E., Var.PIB)
head(PIB)
```

```
##   anyo PIB_M.E. Var.PIB
## 1 2016  289.239     3.3
## 2 2017  280.249     2.5
## 3 2018  275.999     1.4
## 4 2019  255.416     2.1
```

**Inflación:** Es un fenómeno que se observa en la economía de un país y está relacionado con el aumento desordenado de los precios de la mayor parte de los bienes y servicios que se comercian en sus mercados, por un periodo de tiempo prolongado. Datos tomados de aquí

```
# vectores
anyo <- c("2016", "2017", "2018", "2019")
Inflacion <- c(5.75, 4.09, 3.18, 3.80)
#Crear dataframe de vectores
Inflacion <- data.frame(anyo, Inflacion)
head(Inflacion)
```

```
##   anyo Inflacion
## 1 2016      5.75
## 2 2017      4.09
## 3 2018      3.18
## 4 2019      3.80
```

**Desempleo:** Es otra de las variables mas importantes de la macroeconomía, porque afecta directamente el bienestar de las personas. El desempleo es el porcentaje de la fuerza de trabajo que está buscando trabajo activamente y que actualmente se encuentra desempleada. Se toma el promedio anual. Datos tomados de aquí.

```
# vectores
anyo <- c("2016", "2017", "2018", "2019")
Desempleo <- c(9.2, 9.4, 9.7, 10.5)
Var.Desempleo <- c(3.36, 1.99, 3.19, 8.25)
#Crear dataframe de vectores
Desempleo <- data.frame(anyo, Desempleo, Var.Desempleo)
head(Desempleo)
```

```
##   anyo Desempleo Var.Desempleo
## 1 2016      9.2      3.36
## 2 2017      9.4      1.99
## 3 2018      9.7      3.19
## 4 2019     10.5      8.25
```

**Balance fiscal:** Es la diferencia entre ingresos y gastos públicos en un determinado territorio. Datos tomados de aquí

```
# vectores
anyo <- c("2016", "2017", "2018", "2019")
GNC <- c(-4, -3.6, -3.1, -2.5)
#Crear dataframe de vectores
GNC <- data.frame(anyo, GNC)
head(GNC)
```

```
##   anyo  GNC
## 1 2016 -4.0
## 2 2017 -3.6
## 3 2018 -3.1
## 4 2019 -2.5
```

**Balance en cuenta corriente:** Es el conjunto de transacciones de intercambio de bienes y servicios, rentas y transferencias (tanto corrientes como de capital), su saldo determina la capacidad o necesidad de financiación de un país. El dato fue tomado de la cuenta corriente de la balanza de pagos de periodicidad anual. Datos tomados aquí.

```
# vectores
anyo <- c("2016", "2017", "2018", "2019")
Balance_Cuenta_Corriente <- c(-13747.75, -13117.66, -10240.88, -12036.18)
#Crear dataframe de vectores
Balance_Cuenta_Corriente <- data.frame(anyo, Balance_Cuenta_Corriente)
head(Balance_Cuenta_Corriente)
```

```
##   anyo Balance_Cuenta_Corriente
## 1 2016      -13747.75
## 2 2017      -13117.66
## 3 2018      -10240.88
## 4 2019      -12036.18
```

**Tasa de intervención:** Corresponde a la tasa de interés mínima que le cobra el Banco de la República a las entidades financieras por los préstamos que les concede generalmente a un día



y, además, sirve como referencia para establecer la tasa de interés máxima que les paga por recibirles dinero que tengan como excedente. Se toma promedio anual de datos diarias. Datos tomados de aquí.

```
# vectores
anyo <- c("2016", "2017", "2018", "2019")
TIM_promedio<- c(7.10, 6.13, 4.35, 4.25)
#Crear dataframe de vectores
TIM <- data.frame(anyo, TIM_promedio)
head(TIM)
```

```
##   anyo TIM_promedio
## 1 2016          7.10
## 2 2017          6.13
## 3 2018          4.35
## 4 2019          4.25
```

**TRM:** La tasa de cambio representativa del mercado (TRM) es la cantidad de pesos colombianos por un dólar de los Estados Unidos. La TRM se calcula con base en las operaciones de compra y venta de divisas entre intermediarios financieros que transan en el mercado cambiario colombiano, con cumplimiento el mismo día cuando se realiza la negociación de las divisas.

Actualmente la Superintendencia Financiera de Colombia es la que calcula y certifica diariamente la TRM con base en las operaciones registradas el día hábil inmediatamente anterior. Datos tomados de aquí

```
#Se leen los datos -
dataset = read.csv('./data/TRM.csv', check.names = FALSE, encoding = "UTF-8",
                  blank.lines.skip = FALSE, dec=",")

#se conservan unicamente las columnas de año y TRM
df = dataset[1]
df['TRM'] = dataset[3]
df$TRM <- as.numeric(as.character(df$TRM))

#Se agrupa bajo la media
media = df
media = media %>%
  group_by(media[1]) %>%
  summarise(across(.cols = everything(), .fns = mean))

#para la mediana
mediana = df
mediana = mediana %>%
  group_by(mediana[1]) %>%
  summarise(across(.cols = everything(), .fns = median))

#Se genera un dataframe con los datos obtenidos
df = media
colnames(df)[2] <- 'TRM_media'
df['TRM_mediana'] <- mediana[2]
df
```

```
## # A tibble: 4 x 3
```

```
##      Anyo TRM_media TRM_mediana
##      <int>      <dbl>      <dbl>
## 1  2016      3051.      3003.
## 2  2017      2951.      2942.
## 3  2018      2956.      2898.
## 4  2019      3281.      3277.
```

Se unen los datos en un solo dataframe

```
df['PIB_M.E.'] = PIB[2]
df['Var.PIB'] = PIB[3]
df['Inflacion'] = Inflacion[2]
df['Desempleo'] = Desempleo[2]
df['var.Desempleo'] = Desempleo[3]
df['GNC'] = GNC[2]
df['Balance_Cuenta_Corriente'] = Balance_Cuenta_Corriente[2]
df['TIM_promedio'] = TIM[2]
head(df)
```

```
## # A tibble: 4 x 11
##      Anyo TRM_media TRM_mediana PIB_M.E. Var.PIB Inflacion Desempleo var.Desempleo
##      <int>      <dbl>      <dbl>      <dbl>      <dbl>      <dbl>      <dbl>      <dbl>
## 1  2016      3051.      3003.      289.      3.3      5.75      9.2      3.36
## 2  2017      2951.      2942.      280.      2.5      4.09      9.4      1.99
## 3  2018      2956.      2898.      276.      1.4      3.18      9.7      3.19
## 4  2019      3281.      3277.      255.      2.1      3.8      10.5     8.25
## # ... with 3 more variables: GNC <dbl>, Balance_Cuenta_Corriente <dbl>,
## #   TIM_promedio <dbl>
```

**Variables sector minero:** Se adicionan variables relacionadas con el sector minero, para mejorar la efectividad de los modelos. Para cada año se toma el promedio de los datos diarios. Datos tomados de aquí.

```
library(dplyr)
#Se leen los datos -
dataset = read_excel('./data/metales.xlsx')
dataset
```

```
## # A tibble: 3 x 7
##      Anyo compra_Oro compra_Plata compra_Platino venta_Oro venta_Plata
##      <dbl>      <dbl>      <dbl>      <dbl>      <dbl>
## 1  2017    109846.      1294.      86790.    119398.    1617.
## 2  2018    110854.      1191.      80470.    120494.    1489.
## 3  2019    135384.      1370.      87958.    147157.    1712.
## # ... with 1 more variable: venta_Platino <dbl>
df %>% left_join(dataset, by=c("Anyo")) -> df
```

## Capítulo 3. Consolidación de la base

En esta sección se unen las dos bases generadas en las fases anteriores en una sola base

```
library(dplyr)
library(tidyr)
datos_completos%>%mutate(Periodo=as.numeric(as.character(Periodo))) %>%
  inner_join(df, by=c("Periodo" = "Anyo")) %>% replace_na(list(costo_ventas = 0,
    gastos_administracion = 0, otros_gastos=0, costos_financieros=0 ,
    gasto_impuestos=0, ingresos_actividades_ordinarias=0, otros_ingresos = 0,
    ingresos_financieros=0)) -> datos_completos2

datos_completos2 %>% mutate(costos_gastos_totales = gastos_administracion +
  otros_gastos +gasto_impuestos + costo_ventas,
  ingresos_totales = ingresos_actividades_ordinarias +
  ingresos_financieros + otros_ingresos) -> base_modelado

base_modelado <- droplevels(base_modelado)
summary(base_modelado)
```

```
##      NIT                                razon_social    CIIU
## Length:82          ANGLOGOLD ASHANTI COLOMBIA S.A.      : 3    B0722:82
## Class :character    CALDAS GOLD MARMATO S.A.S.          : 3
## Mode  :character    CONTINENTAL GOLD LIMITED SUCURSAL COLOMBIA: 3
##                                ECO ORO MINERALS CORP      : 3
##                                EXPLORACIONES CHAPARRAL COLOMBIA SAS : 3
##                                EXPLORACIONES NORTHERN COLOMBIA S.A.S : 3
##                                (Other)                    :64
##      ciudad      departamento      estado      situacion
## BOGOTÁ, D.C.:26    ANTIOQUIA      :44    INSPECCION:41    ACTIVA:82
## BUCARAMANGA :12    BOGOTÁ, D. C.:26    VIGILANCIA:41
## ENVIGADO      : 2    SANTANDER      :12
## MEDELLÍN      :42
##
##
##      organo_societario etapa_situacion      Periodo
## ACTIVIDAD ECONOMICA DIFERENTE:82      ACTIVA:82      Min.      :2017
##                                           1st Qu.:2017
##                                           Median :2018
##                                           Mean   :2018
##                                           3rd Qu.:2019
```

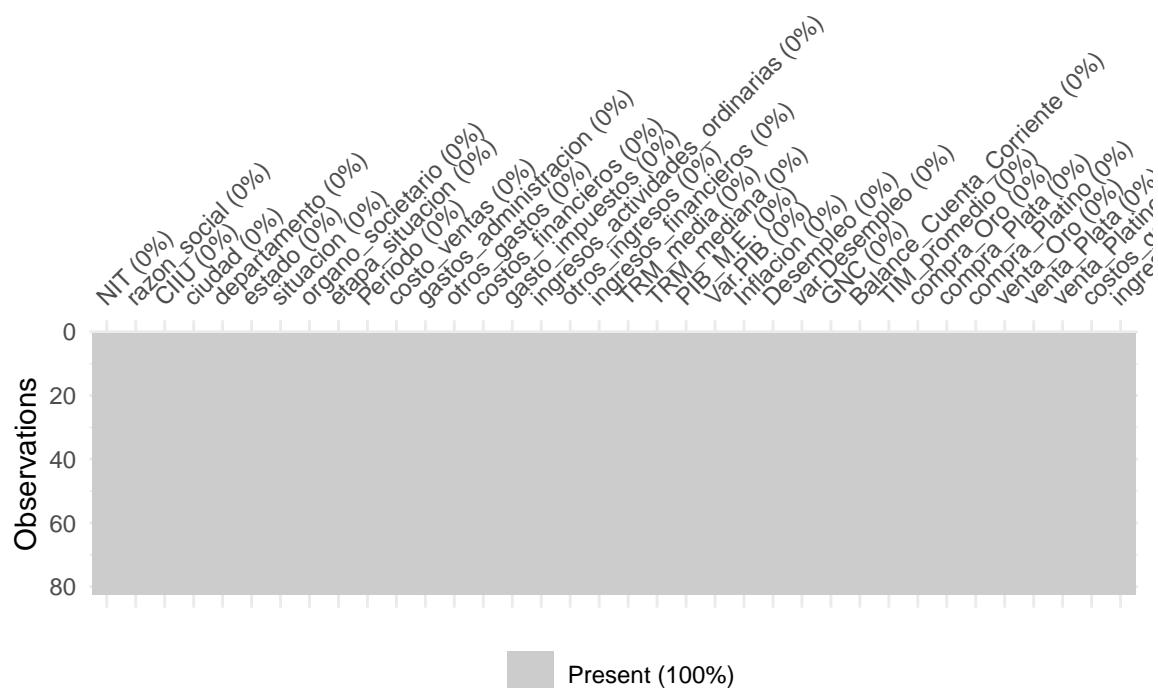
```

##                                     Max.      :2019
##
## costo_ventas      gastos_administracion  otros_gastos
## Min.      :      0    Min.      :      0    Min.      :      0
## 1st Qu.:      0    1st Qu.:    15366    1st Qu.:     848
## Median :      0    Median :    560711    Median :   120317
## Mean   : 23244524    Mean   : 12272119    Mean   : 3275946
## 3rd Qu.: 6666040    3rd Qu.: 5672411    3rd Qu.: 1736150
## Max.    :408474390    Max.    :287863704    Max.    :46689684
##
## costos_financieros gasto_impuestos      ingresos_actividades_ordinarias
## Min.      :      0    Min.      : -8974634    Min.      :      0
## 1st Qu.:      0    1st Qu.:      0    1st Qu.:      0
## Median : 168546    Median :     2546    Median :      0
## Mean   : 2983081    Mean   : 4495562    Mean   : 34205451
## 3rd Qu.: 2383609    3rd Qu.: 141459    3rd Qu.:      0
## Max.    :52689630    Max.    :131684824    Max.    :954650443
##
## otros_ingresos      ingresos_financieros      TRM_media      TRM_mediana
## Min.      :      0    Min.      :      0    Min.      :2951    Min.      :2898
## 1st Qu.:      0    1st Qu.:      0    1st Qu.:2951    1st Qu.:2898
## Median : 36549    Median :     3942    Median :2956    Median :2942
## Mean   : 836610    Mean   : 1285604    Mean   :3058    Mean   :3033
## 3rd Qu.: 601104    3rd Qu.: 850632    3rd Qu.:3281    3rd Qu.:3277
## Max.    :11404344    Max.    :26100695    Max.    :3281    Max.    :3277
##
## PIB_M.E.      Var.PIB      Inflacion      Desempleo
## Min.      :255.4    Min.      :1.400    Min.      :3.180    Min.      : 9.400
## 1st Qu.:255.4    1st Qu.:1.400    1st Qu.:3.180    1st Qu.: 9.400
## Median :276.0    Median :2.100    Median :3.800    Median : 9.700
## Mean   :270.9    Mean   :1.998    Mean   :3.687    Mean   : 9.851
## 3rd Qu.:280.2    3rd Qu.:2.500    3rd Qu.:4.090    3rd Qu.:10.500
## Max.    :280.2    Max.    :2.500    Max.    :4.090    Max.    :10.500
##
## var.Desempleo      GNC      Balance_Cuenta_Corriente      TIM_promedio
## Min.      :1.990    Min.      : -3.60    Min.      : -13118    Min.      :4.250
## 1st Qu.:1.990    1st Qu.: -3.60    1st Qu.: -13118    1st Qu.:4.250
## Median :3.190    Median : -3.10    Median : -12036    Median :4.350
## Mean   :4.385    Mean   : -3.08    Mean   : -11792    Mean   :4.926
## 3rd Qu.:8.250    3rd Qu.: -2.50    3rd Qu.: -10241    3rd Qu.:6.130
## Max.    :8.250    Max.    : -2.50    Max.    : -10241    Max.    :6.130
##
## compra_Oro      compra_Plata      compra_Platino      venta_Oro
## Min.      :109846    Min.      :1191    Min.      :80470    Min.      :119398
## 1st Qu.:109846    1st Qu.:1191    1st Qu.:80470    1st Qu.:119398
## Median :110854    Median :1294    Median :86790    Median :120494
## Mean   :118288    Mean   :1283    Mean   :85002    Mean   :128574
## 3rd Qu.:135384    3rd Qu.:1370    3rd Qu.:87958    3rd Qu.:147157
## Max.    :135384    Max.    :1370    Max.    :87958    Max.    :147157
##
## venta_Plata      venta_Platino      costos_gastos_totales      ingresos_totales

```

```
## Min.      :1489    Min.      :83441    Min.      :    2869    Min.      :    0
## 1st Qu.:1489    1st Qu.:83441    1st Qu.:   528039    1st Qu.:   24364
## Median :1617    Median :89994    Median :  4175588    Median :   587346
## Mean    :1604    Mean    :88140    Mean    : 43288150    Mean    : 36327665
## 3rd Qu.:1712    3rd Qu.:91205    3rd Qu.: 26972121    3rd Qu.: 3698197
## Max.     :1712    Max.     :91205    Max.     :700560590    Max.     :967119965
##
```

```
library(visdat)
vis_miss(base_modelado)
```



Esta gráfica nos ayuda a visualizar que no hay datos perdidos en la data. # Capítulo 4. Análisis descriptivo

A continuación se realiza un análisis descriptivo de los datos

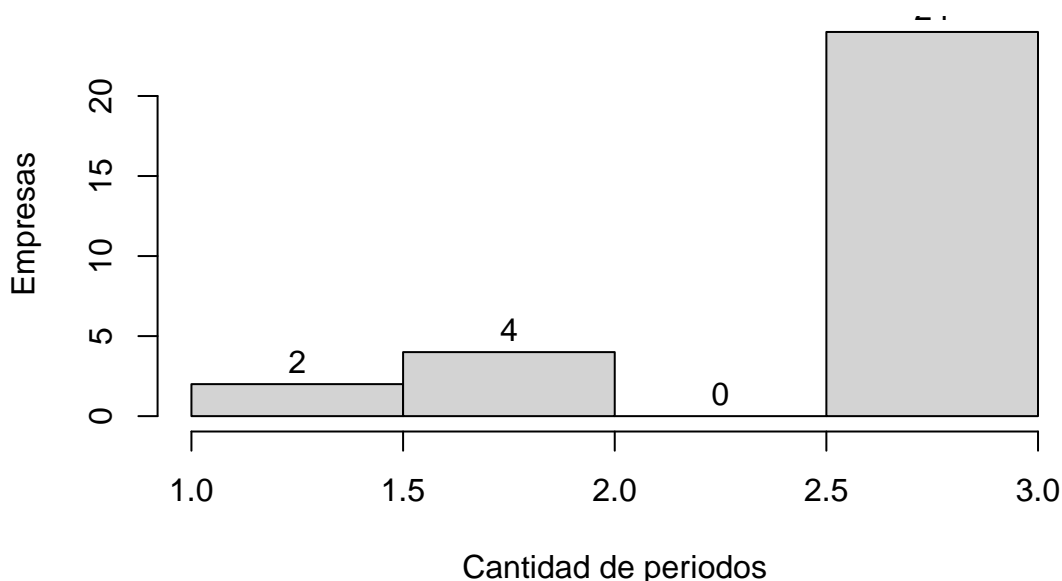
```
library(ggplot2)
library(tidyr)
library(dplyr)

datosValidar <- base_modelado %>% group_by(NIT, razon_social) %>%
  summarise(total=n())

h <- hist(x = datosValidar$total, main = "Cantidad de periodos por empresa",
          xlab = "Cantidad de periodos", ylab = "Empresas", breaks=3)

text(h$mids,h$counts,labels=h$counts, adj=c(0.5, -0.5))
```

### Cantidad de periodos por empresa



Podemos observar que hay empresas que no tienen los 3 periodos, solo trabajaremos con la empresas que tengan los periodos completos.

```
base_modelado %>% anti_join(datosValidar %>% filter(total < 3) , by="NIT" ) %>%
  group_by(NIT, razon_social) %>% summarise(total=n()) %>% arrange(desc(total))
```

```
## # A tibble: 24 x 3
## # Groups:   NIT [24]
##   NIT      razon_social      total
##   <chr>    <fct>          <int>
## 1 811002172 MINERA CROESUS S.A.S      3
## 2 830012565 ECO ORO MINERALS CORP    3
## 3 830127076 ANGLOGOLD ASHANTI COLOMBIA S.A. 3
## 4 860507991 SANTIAGO OIL COMPANY      3
## 5 890114642 CALDAS GOLD MARMATO S.A.S. 3
## 6 900039998 MINERALES ANDINOS DE OCCIDENTE S.A 3
## 7 900062755 MINERIA INTEGRAL DE COLOMBIA S.A.S. 3
## 8 900063262 SOCIEDAD MINERA DE SANTANDER S.A.S. 3
## 9 900084407 GRAMALOTE COLOMBIA LIMITED    3
## 10 900156833 MINERA DE COBRE QUEBRADONA SA    3
## # ... with 14 more rows
```

```
base_modelado %>% anti_join(datosValidar %>% filter(total < 3) , by="NIT" ) ->
  base_modelado
base_modelado %>% filter(ingresos_totales == 0) %>% count(NIT) %>%
  filter(n==3) -> datosValidar
base_modelado %>% anti_join(datosValidar, by="NIT" ) -> base_modelado

datosValidar <- base_modelado %>% group_by(NIT, razon_social) %>%
  summarise(total=n())
```

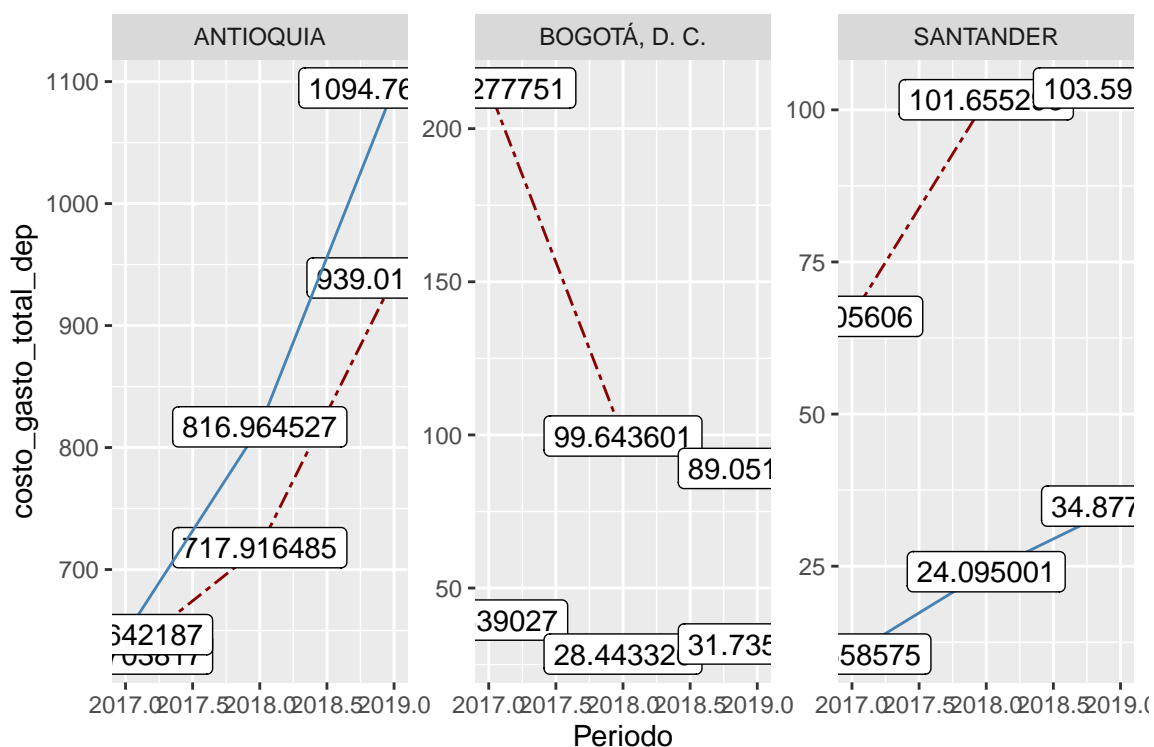
```
table(datosValidar$total)
```

```
##
## 3
## 23
```

Ya podemos ver que tenemos 23 empresas con los 3 periodos. Veamos los ingresos y los costos por departamento.

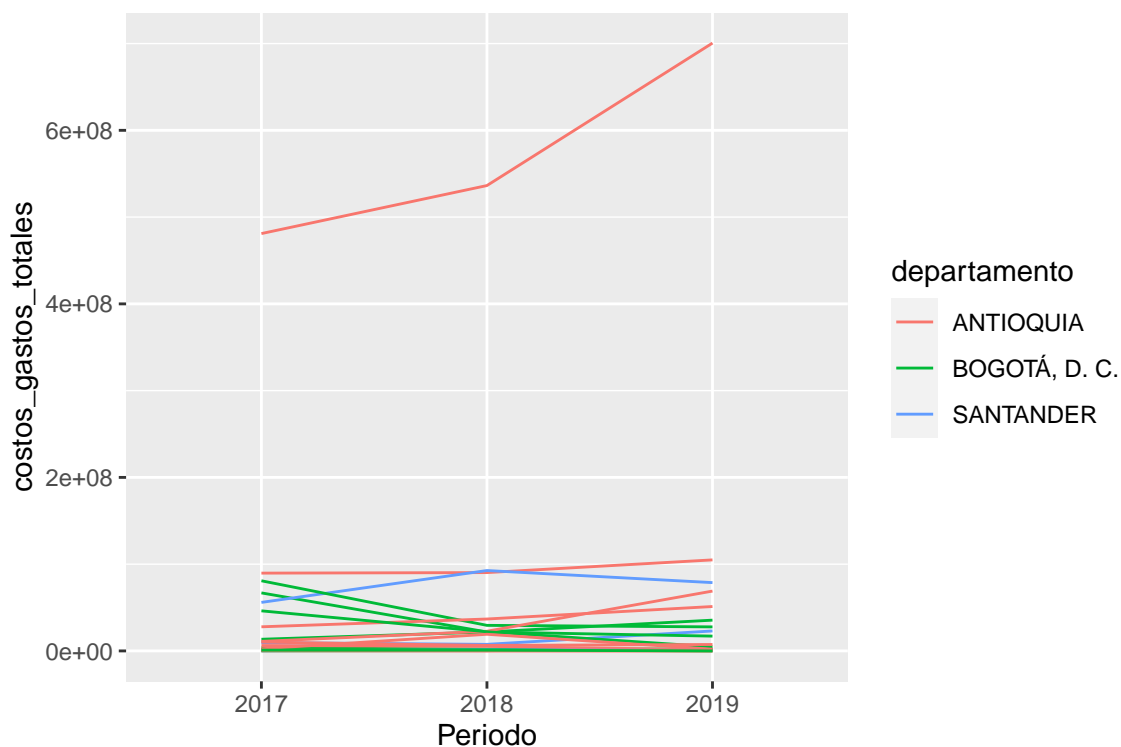
```
datosValidarDepartamento <- base_modelado %>% group_by(departamento, Periodo) %>%
  summarise(costo_gasto_total_dep = sum(costos_gastos_totales) / 1000000,
    ingresos_totales_dep = sum(ingresos_totales) / 1000000)
```

```
ggplot(datosValidarDepartamento, aes(x= Periodo))+
  geom_line(aes(y = costo_gasto_total_dep, color="darkred", linetype="twodash"))+
  geom_label(aes(y = costo_gasto_total_dep, label=costo_gasto_total_dep)) +
  geom_line(aes(y = ingresos_totales_dep, label="Ingresos"), color = "steelblue")+
  geom_label(aes(y = ingresos_totales_dep, label=ingresos_totales_dep)) +
  facet_wrap(~departamento, scales = "free_y")
```



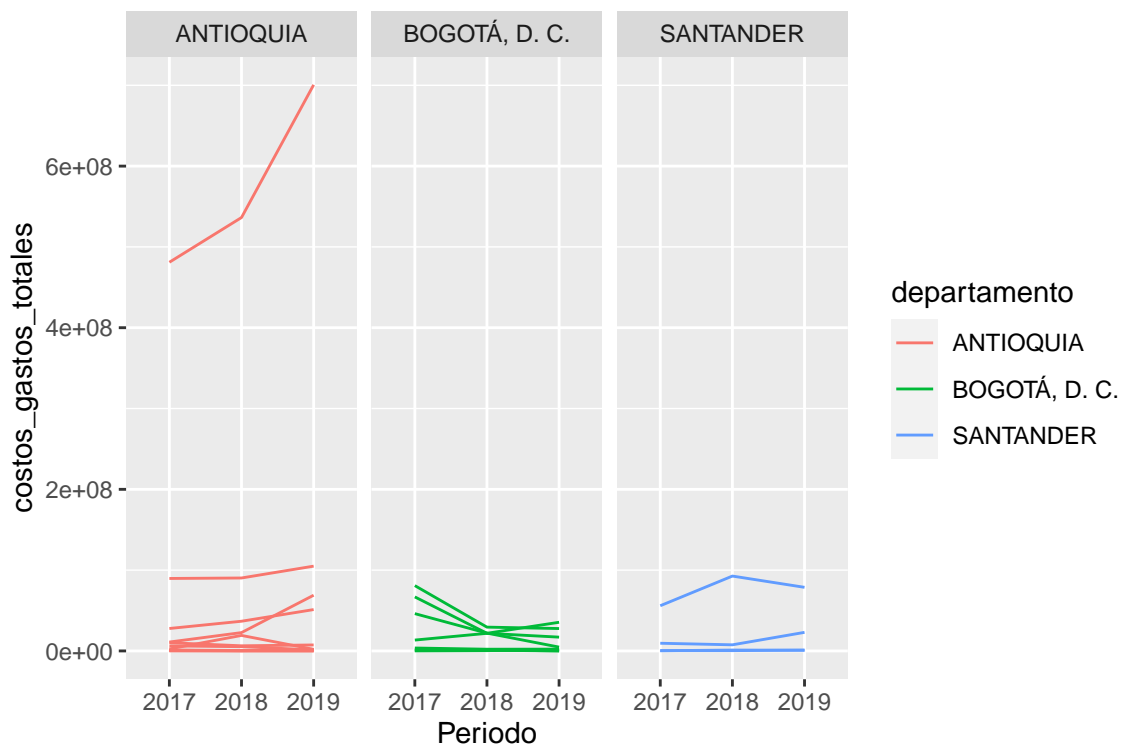
```
base_modelado$NIT=as.factor(base_modelado$NIT)
base_modelado$Periodo=as.factor(base_modelado$Periodo)
```

```
p1=ggplot(base_modelado, aes(y=costos_gastos_totales,x=Periodo,group=NIT,colour=departamento))
p1+geom_line()
```



Se pueden ver algunos comportamientos diferentes por departamento, sin embargo separemos el gráfico para ver mejor:

```
p1+geom_line()+facet_grid(~departamento)
```

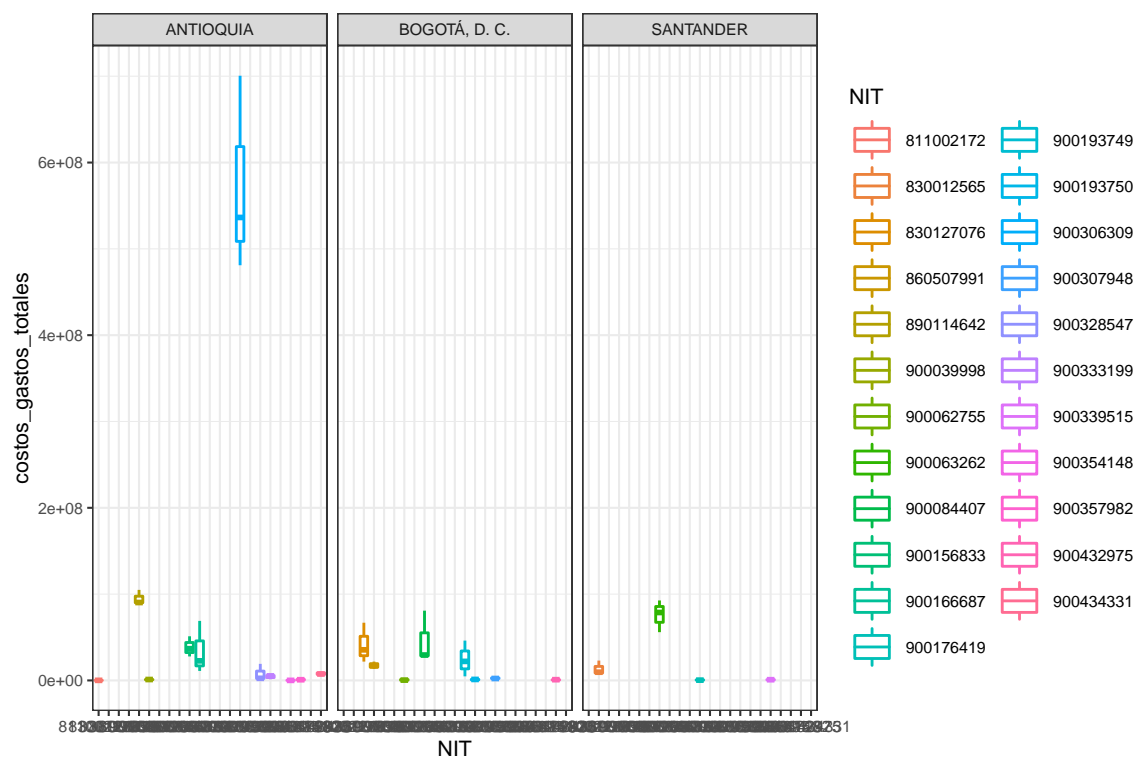


El gráfico anterior nos muestra que cada empresa tiene costos/gastos totales particulares. Adicionalmente, hay una empresa de Medellín que tiene costos/gastos totales mas altos, comparada con las otras. Tratemos de identificar las empresas que tienen un comportamiento más diferente a las demás.



```
theme_set(theme_bw(base_size = 8))

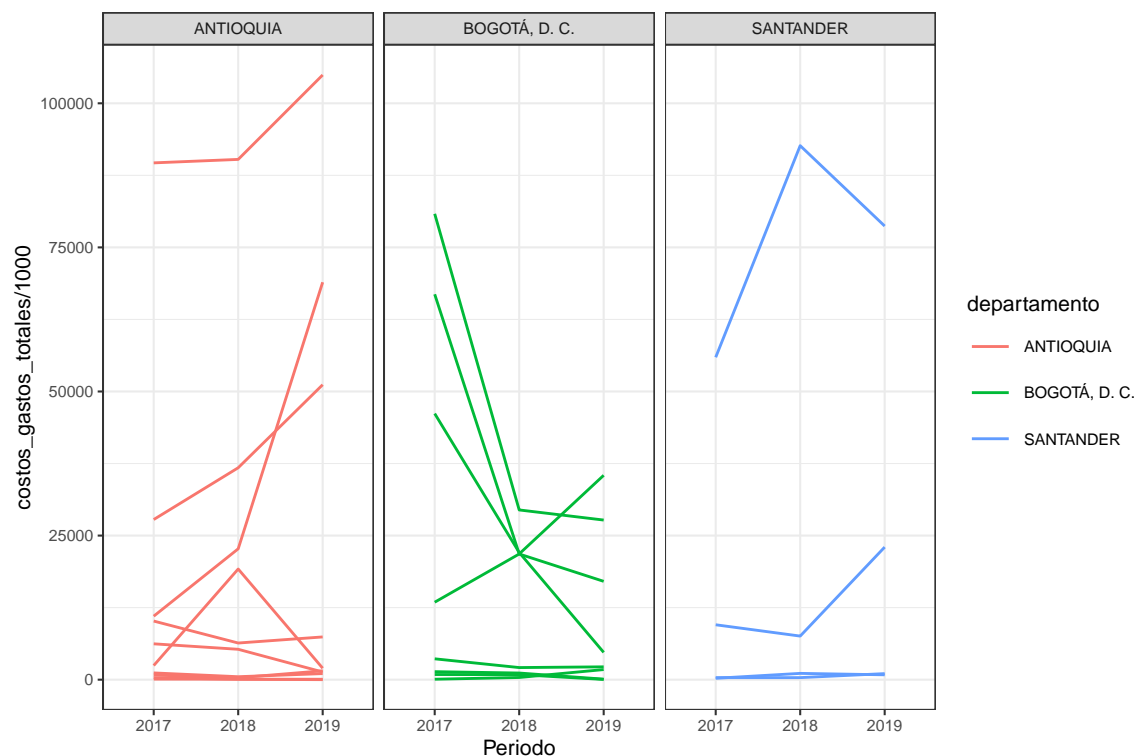
qplot(NIT, costos_gastos_totales, facets = . ~ departamento,
      colour = NIT, geom = "boxplot", data = base_modelado)
```



Al

parecer solo hay 1 empresa que tiene comportamiento de costos/gastos totales mucho mas diferente a las demás. Realizaremos el ejercicio de eliminar (solo para efectos visuales) la empresa que es mas diferente a las demas.

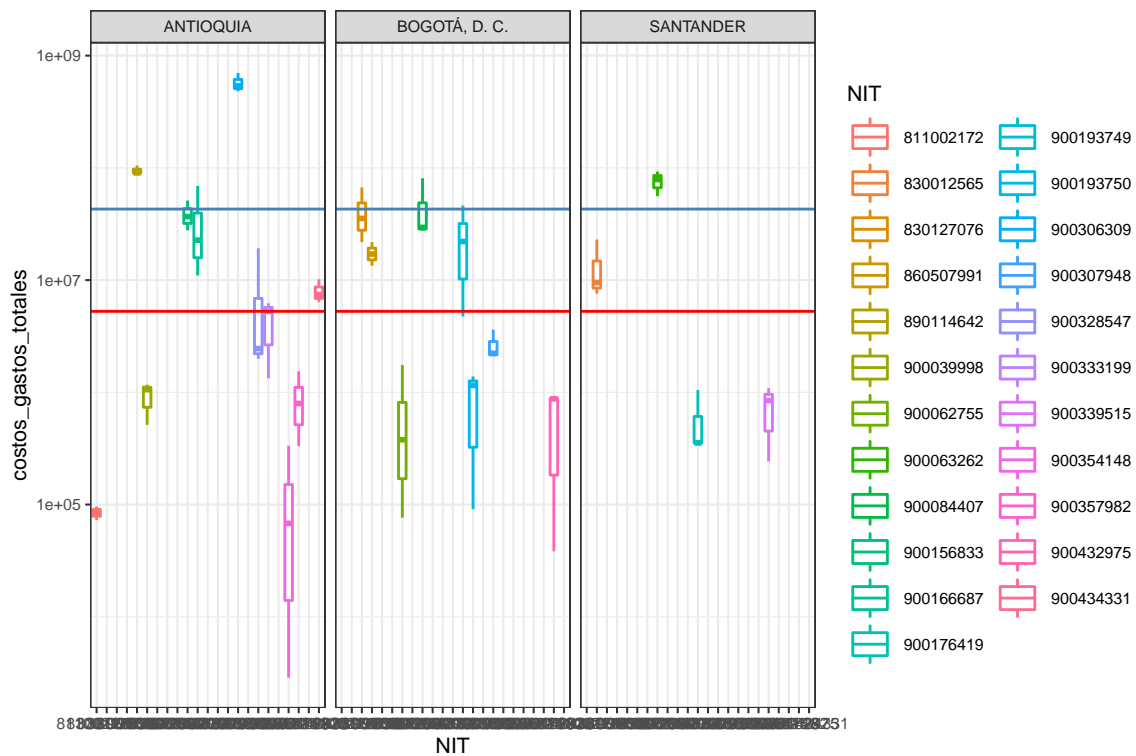
```
datosValidar <- filter(base_modelado, NIT!=900306309)
p1=ggplot(datosValidar, aes(y=costos_gastos_totales/1000,x=Periodo,group=NIT,
                           colour=departamento))
p1+geom_line()+facet_grid(.~departamento)
```



Confirmamos que los costos/gastos totales son particulares de cada empresa. Cambiemos la escala de los datos y volvamos a graficar, para poder apreciar mejor el comportamiento de las otras empresas que tienen costos/gastos totales mas bajos, pero con el set de empresas completo.

```
theme_set(theme_bw(base_size = 8))

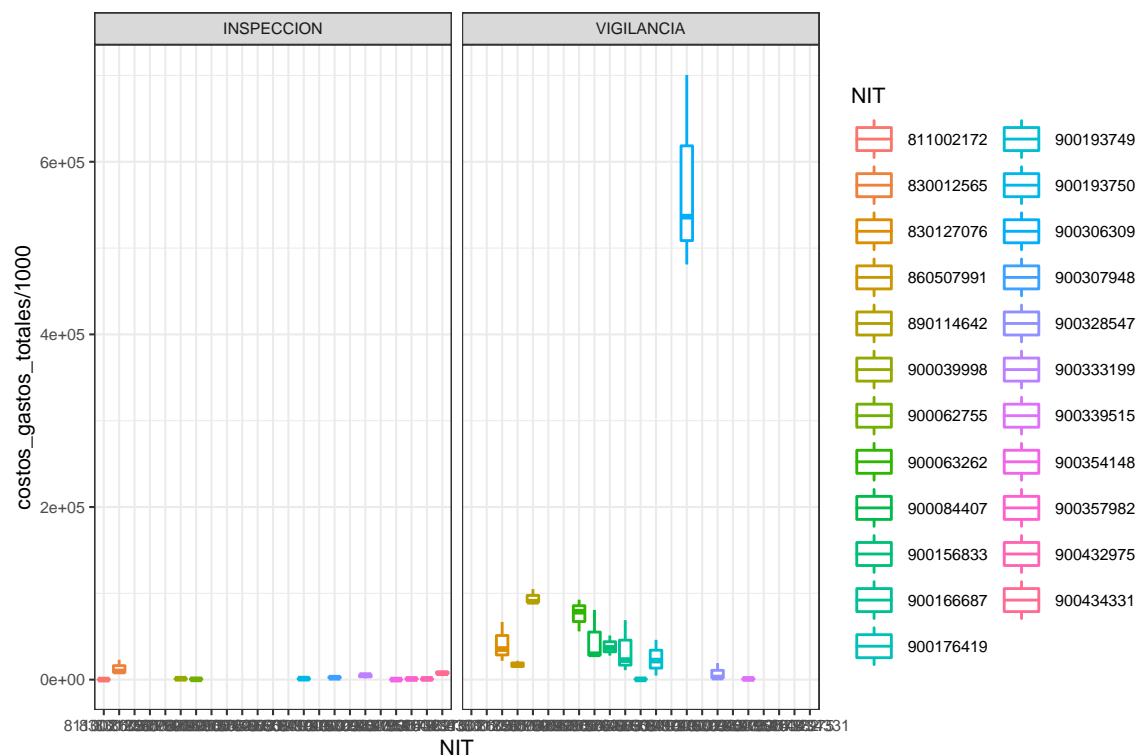
qplot(NIT, costos_gastos_totales, facets = . ~ departamento,
      colour = NIT, geom = "boxplot", data = base_modelado) +
  scale_y_log10() +
  geom_hline(aes(yintercept = mean(costos_gastos_totales)), color = "steelblue") +
  geom_hline(aes(yintercept = median(costos_gastos_totales)), color = "red")
```



Ahora podemos ver mejor que cada empresa tiene unos costos/gastos totales particulares, así como costos promedio diferentes. Además, encontramos que solamente hay 5 empresas que tienen un comportamiento general en sus costos/gastos totales. Ahora revisemos los costos/gastos totales con el estado.

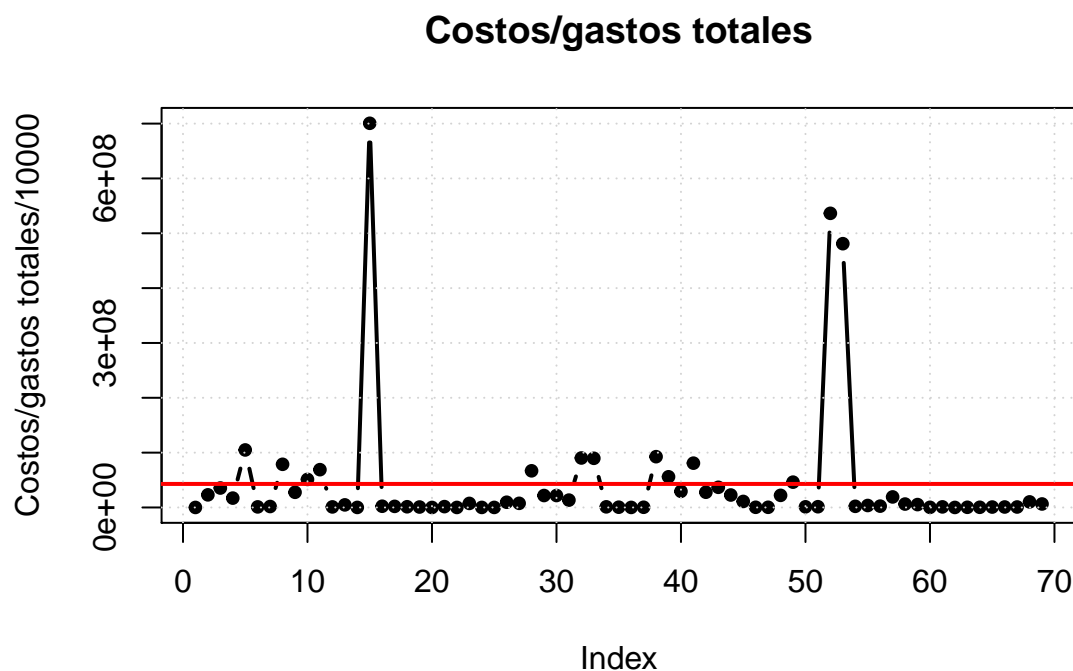
```
theme_set(theme_bw(base_size = 8))

qplot(NIT, costos_gastos_totales/1000, facets = . ~ estado,
      colour = NIT, geom = "boxplot", data = base_modelado)
```



Podemos ver que las empresas con estado inspección presentan costos/gastos totales menores que las empresas con estado vigilancia. Veamos ahora la dispersion de nuestra variable objetivo.

```
plot(base_modelado$costos_gastos_totales, main="Costos/gastos totales",
      type="b", ylab="Costos/gastos totales/10000", pch= 20, lwd=2)
abline(h=mean(base_modelado$costos_gastos_totales), lwd=2, col= "red")
grid()
```



Con esto confirmamos que la dispersion de los costos/gastos totales no tiene un compor-

tamiento general.

## Estandarización de variables

A continuación se realiza un reescalamiento de las variables para que sean comparables y así apuntarle a mejorar la calidad de los modelos

```
library(scales)
library(kableExtra)
library(tibble)
base.escalada<- scale(base_modelado[,c(11,12, 13,14,15,16,17,18,19,20,21,22,23,
24,25,26,27,28,29,30,31,32,33,34,35,36)],center=T,scale=T)
base.escalada<- as.data.frame(base.escalada)
base.escalada<- rownames_to_column(base.escalada)
tmp <- rownames_to_column(base_modelado %>% select(c("NIT", "razon_social",
"CIIU", "ciudad","departamento", "estado", "situacion", "organo_societario",
"etapa_situacion", "Periodo")))
base_modelado <- tmp %>% left_join(base.escalada, by="rowname") %>%
select(-c("rowname"))
```



# Capítulo 5. Correlaciones

En este capítulo realizaremos un análisis de las correlaciones entre las variables que nos ayudarán a decidir que variables pueden ser las mejores candidatas para nuestro modelo.

La base de datos que utilizaremos es `base_modelado` la misma que visualizamos en la página 15

```
head(base_modelado)
```

```
##          NIT                                razon_social CIIU          ciudad departamento
## 1 811002172                                MINERA CROESUS S.A.S B0722    MEDELLÍN    ANTIOQUIA
## 2 830012565                                ECO ORO MINERALS CORP B0722  BUCARAMANGA    SANTANDER
## 3 830127076    ANGLOGOLD ASHANTI COLOMBIA S.A. B0722  BOGOTÁ, D.C. BOGOTÁ, D. C.
## 4 860507991                                SANTIAGO OIL COMPANY B0722  BOGOTÁ, D.C. BOGOTÁ, D. C.
## 5 890114642                                CALDAS GOLD MARMATO S.A.S. B0722    MEDELLÍN    ANTIOQUIA
## 6 900039998  MINERALES ANDINOS DE OCCIDENTE S.A B0722    MEDELLÍN    ANTIOQUIA
##          estado situacion                                organo_societario etapa_situacion Periodo
## 1 INSPECCION    ACTIVA ACTIVIDAD ECONOMICA DIFERENTE                                ACTIVA    2019
## 2 INSPECCION    ACTIVA ACTIVIDAD ECONOMICA DIFERENTE                                ACTIVA    2019
## 3 VIGILANCIA    ACTIVA ACTIVIDAD ECONOMICA DIFERENTE                                ACTIVA    2019
## 4 VIGILANCIA    ACTIVA ACTIVIDAD ECONOMICA DIFERENTE                                ACTIVA    2019
## 5 VIGILANCIA    ACTIVA ACTIVIDAD ECONOMICA DIFERENTE                                ACTIVA    2019
## 6 INSPECCION    ACTIVA ACTIVIDAD ECONOMICA DIFERENTE                                ACTIVA    2019
##          costo_ventas gastos_administracion otros_gastos costos_financieros
## 1 -0.33867199          -0.4074151    -0.4014073          -0.01766357
## 2 -0.33867199          -0.1245786     1.7804892          -0.41681165
## 3  0.14621874          -0.4096184    -0.4014073          -0.11939539
## 4 -0.04073369          -0.3719525    -0.3443104           0.18991196
## 5  0.96663837          -0.3124819    -0.3166147          -0.12092215
## 6 -0.33867199          -0.4096184    -0.2576751           0.45676213
##          gasto_impuestos ingresos_actividades_ordinarias otros_ingresos
## 1 -0.22397057          -0.23958426     0.005953155
## 2 -0.22262585          -0.23958426    -0.210004581
## 3 -0.22477997          -0.23958426    -0.508192736
## 4 -0.48682468          -0.06443343    -0.363463888
## 5  0.05505802           0.51657397     0.383638219
## 6 -0.22467562          -0.23958426    -0.508192736
##          ingresos_financieros TRM_media TRM_mediana PIB_M.E.    Var.PIB Inflacion
## 1 -0.3515285     1.4038     1.39604 -1.385842 0.2183709 0.2877006
## 2 -0.3515285     1.4038     1.39604 -1.385842 0.2183709 0.2877006
## 3 -0.3515285     1.4038     1.39604 -1.385842 0.2183709 0.2877006
## 4 -0.1320566     1.4038     1.39604 -1.385842 0.2183709 0.2877006
## 5  0.4181071     1.4038     1.39604 -1.385842 0.2183709 0.2877006
## 6 -0.3515285     1.4038     1.39604 -1.385842 0.2183709 0.2877006
```

```

##      Desempleo var.Desempleo      GNC Balance_Cuenta_Corriente TIM_promedio
## 1  1.354199      1.380845 1.250959      -0.1990925 -0.7586531
## 2  1.354199      1.380845 1.250959      -0.1990925 -0.7586531
## 3  1.354199      1.380845 1.250959      -0.1990925 -0.7586531
## 4  1.354199      1.380845 1.250959      -0.1990925 -0.7586531
## 5  1.354199      1.380845 1.250959      -0.1990925 -0.7586531
## 6  1.354199      1.380845 1.250959      -0.1990925 -0.7586531
##      compra_Oro compra_Plata compra_Platino venta_Oro venta_Plata venta_Platino
## 1  1.403075      1.151478      0.8708067 1.403075      1.15238      0.8708068
## 2  1.403075      1.151478      0.8708067 1.403075      1.15238      0.8708068
## 3  1.403075      1.151478      0.8708067 1.403075      1.15238      0.8708068
## 4  1.403075      1.151478      0.8708067 1.403075      1.15238      0.8708068
## 5  1.403075      1.151478      0.8708067 1.403075      1.15238      0.8708068
## 6  1.403075      1.151478      0.8708067 1.403075      1.15238      0.8708068
##      costos_gastos_totales ingresos_totales
## 1      -0.36104643      -0.24581787
## 2      -0.16782320      -0.24842906
## 3      -0.06283635      -0.25203452
## 4      -0.21786041      -0.07148885
## 5       0.52255596       0.52606828
## 6      -0.35276218      -0.25203452

```

Para el análisis de correlaciones, se toman las variables macroeconómicas (PIB, Inflación, Desempleo, GNC, Balance de cuenta corriente y TRM), compra y venta de metales preciosos y se compara con la variable objetivo.

### Calcular el coeficiente de correlación

Este comando calcula la matriz de correlación:

```

##      TRM_media TRM_mediana PIB_M.E. Var.PIB Inflacion
## TRM_media      1.00      0.99      -0.99      0.14      0.19
## TRM_mediana      0.99      1.00      -0.96      0.26      0.31
## PIB_M.E.      -0.99      -0.96      1.00      0.00      -0.05
## Var.PIB      0.14      0.26      0.00      1.00      1.00
## Inflacion      0.19      0.31      -0.05      1.00      1.00
## Desempleo      0.97      0.93      -0.99      -0.11      -0.06
## var.Desempleo      0.99      0.96      -1.00      -0.03      0.02
## GNC      0.90      0.84      -0.95      -0.31      -0.26
## Balance_Cuenta_Corriente      -0.13      -0.25      -0.02      -1.00      -1.00
## TIM_promedio      -0.55      -0.45      0.67      0.75      0.71
## compra_Oro      1.00      0.99      -0.99      0.12      0.17
## compra_Plata      0.81      0.88      -0.72      0.69      0.73
## compra_Platino      0.61      0.70      -0.49      0.87      0.89
## venta_Oro      1.00      0.99      -0.99      0.12      0.17
## venta_Plata      0.81      0.88      -0.72      0.69      0.73
## venta_Platino      0.61      0.70      -0.49      0.87      0.89
## Var.objetivo      0.04      0.04      -0.04      0.00      0.01
##      Desempleo var.Desempleo      GNC Balance_Cuenta_Corriente
## TRM_media      0.97      0.99      0.90      -0.13
## TRM_mediana      0.93      0.96      0.84      -0.25
## PIB_M.E.      -0.99      -1.00      -0.95      -0.02

```



## Var.PIB	-0.11	-0.03	-0.31	-1.00
## Inflacion	-0.06	0.02	-0.26	-1.00
## Desempleo	1.00	1.00	0.98	0.12
## var.Desempleo	1.00	1.00	0.96	0.04
## GNC	0.98	0.96	1.00	0.32
## Balance_Cuenta_Corriente	0.12	0.04	0.32	1.00
## TIM_promedio	-0.74	-0.68	-0.86	-0.76
## compra_Oro	0.97	0.99	0.91	-0.11
## compra_Plata	0.64	0.70	0.47	-0.68
## compra_Platino	0.39	0.47	0.20	-0.86
## venta_Oro	0.97	0.99	0.91	-0.11
## venta_Plata	0.64	0.70	0.47	-0.68
## venta_Platino	0.39	0.47	0.20	-0.86
## Var.objetivo	0.04	0.04	0.03	0.00
##	TIM_promedio	compra_Oro	compra_Plata	compra_Platino
## TRM_media	-0.55	1.00	0.81	0.61
## TRM_mediana	-0.45	0.99	0.88	0.70
## PIB_M.E.	0.67	-0.99	-0.72	-0.49
## Var.PIB	0.75	0.12	0.69	0.87
## Inflacion	0.71	0.17	0.73	0.89
## Desempleo	-0.74	0.97	0.64	0.39
## var.Desempleo	-0.68	0.99	0.70	0.47
## GNC	-0.86	0.91	0.47	0.20
## Balance_Cuenta_Corriente	-0.76	-0.11	-0.68	-0.86
## TIM_promedio	1.00	-0.57	0.04	0.32
## compra_Oro	-0.57	1.00	0.80	0.59
## compra_Plata	0.04	0.80	1.00	0.96
## compra_Platino	0.32	0.59	0.96	1.00
## venta_Oro	-0.57	1.00	0.80	0.59
## venta_Plata	0.04	0.80	1.00	0.96
## venta_Platino	0.32	0.59	0.96	1.00
## Var.objetivo	-0.02	0.04	0.03	0.02
##	venta_Oro	venta_Plata	venta_Platino	Var.objetivo
## TRM_media	1.00	0.81	0.61	0.04
## TRM_mediana	0.99	0.88	0.70	0.04
## PIB_M.E.	-0.99	-0.72	-0.49	-0.04
## Var.PIB	0.12	0.69	0.87	0.00
## Inflacion	0.17	0.73	0.89	0.01
## Desempleo	0.97	0.64	0.39	0.04
## var.Desempleo	0.99	0.70	0.47	0.04
## GNC	0.91	0.47	0.20	0.03
## Balance_Cuenta_Corriente	-0.11	-0.68	-0.86	0.00
## TIM_promedio	-0.57	0.04	0.32	-0.02
## compra_Oro	1.00	0.80	0.59	0.04
## compra_Plata	0.80	1.00	0.96	0.03
## compra_Platino	0.59	0.96	1.00	0.02
## venta_Oro	1.00	0.80	0.59	0.04
## venta_Plata	0.80	1.00	0.96	0.03
## venta_Platino	0.59	0.96	1.00	0.02
## Var.objetivo	0.04	0.03	0.02	1.00

Con el resultado de la correlación, podemos interpretar que las variables macroeconomicas,

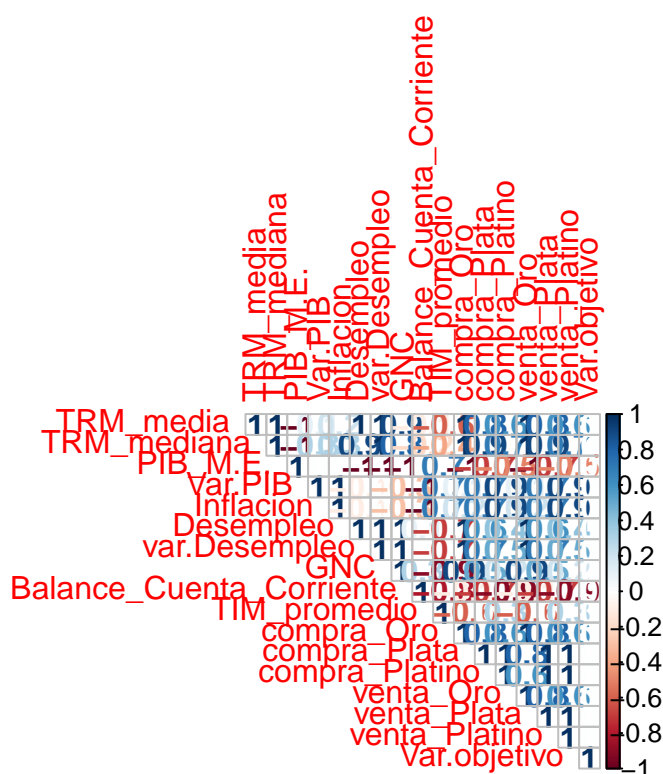
compra y venta de metales preciosos no explican la variable objetivo, ya que el nivel de significancia es cercano a cero. Es decir, no encontramos una asociación entre estas variables y la variable objetivo, que nos ayude a predecir o explicar el comportamiento de los costos y gastos totales de las empresas del sector “extracción de oro y otros metales preciosos”.

## Visualizar matriz de correlación

Utilizaremos el comando `corrplot` para visualizar la matriz de correlaciones. Lo primero es calcular la matriz, guardarla en un objeto y luego graficarla. En este caso vamos a graficar los coeficientes resultantes.

```
correlacion<-round(cor(base), 1)

corrplot(correlacion, method="number", type="upper")
```

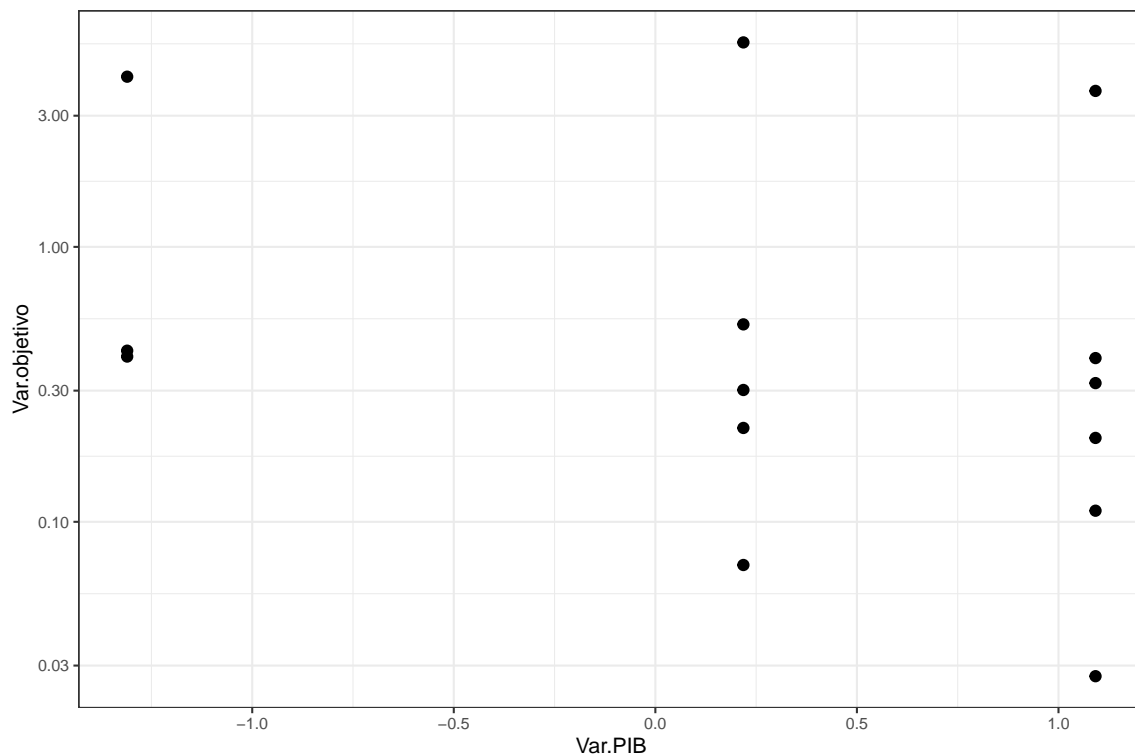


Se

realizó el análisis de correlación con el estadístico de Pearson, con las variables originales y estandarizadas, con el fin de evidenciar las variables explicativas de la variable objetivo construida, encontrando como resultado la no significancia para explicar el comportamiento de los costos y gastos del sector que estamos estudiando.

A continuación se grafican los datos de la Variable objetivo (costos y gastos totales) con respecto a var.PIB (Variación del PIB)

```
ggplot(base, aes(x=Var.PIB, y=Var.objetivo)) +geom_point()+scale_y_log10()
```



Podemos apreciar que no se presenta un comportamiento lineal entre las variables, por lo tanto la variable PIB no es una buena candidata para utilizar en nuestro modelo. De forma general encontramos que las macroeconómicas no explican la variable objetivo. Por tal motivo se decide no trabajar con las variables macroeconomicas venta y compra de metales preciosos en los modelos.

Sin embargo adicionamos a nuestra base de datos estas variables económicas que servirán de insumo para la fase de modelado.

```
base_modelado %>% select(departamento, estado, Periodo, costos_gastos_totales,
ciudad, ingresos_totales, NIT, razon_social, otros_ingresos ,
ingresos_financieros, compra_Oro, compra_Plata, compra_Platino, venta_Oro,
venta_Plata, venta_Platino) -> base_modelo_lineal
```

*#definir variable de tamaño de la empresa*

```
bussiness_size <- cut(base_modelo_lineal$costos_gastos_totales, breaks=4)
levels(bussiness_size) <- list(small = "(-0.377,0.349]",
medium = "(0.349,1.07]",
big = "(1.07,1.8]",
very_big="(1.8,2.52]")
base_modelo_lineal['tamano_empresa'] <- bussiness_size
base_modelo_lineal %>% filter(NIT != 900306309) -> base_modelo_lineal
head(base_modelo_lineal)
```

```
##      departamento      estado Periodo costos_gastos_totales      ciudad
## 1      ANTIOQUIA INSPECCION    2019      -0.36104643      MEDELLÍN
## 2      SANTANDER INSPECCION    2019      -0.16782320      BUCARAMANGA
## 3 BOGOTÁ, D. C. VIGILANCIA    2019      -0.06283635      BOGOTÁ, D.C.
## 4 BOGOTÁ, D. C. VIGILANCIA    2019      -0.21786041      BOGOTÁ, D.C.
## 5      ANTIOQUIA VIGILANCIA    2019       0.52255596      MEDELLÍN
## 6      ANTIOQUIA INSPECCION    2019      -0.35276218      MEDELLÍN
```

##	ingresos_totales	NIT	razon_social	otros_ingresos
## 1	-0.24581787	811002172	MINERA CROESUS S.A.S	0.005953155
## 2	-0.24842906	830012565	ECO ORO MINERALS CORP	-0.210004581
## 3	-0.25203452	830127076	ANGLOGOLD ASHANTI COLOMBIA S.A.	-0.508192736
## 4	-0.07148885	860507991	SANTIAGO OIL COMPANY	-0.363463888
## 5	0.52606828	890114642	CALDAS GOLD MARMATO S.A.S.	0.383638219
## 6	-0.25203452	900039998	MINERALES ANDINOS DE OCCIDENTE S.A	-0.508192736

##	ingresos_financieros	compra_Oro	compra_Plata	compra_Platino	venta_Oro
## 1	-0.3515285	1.403075	1.151478	0.8708067	1.403075
## 2	-0.3515285	1.403075	1.151478	0.8708067	1.403075
## 3	-0.3515285	1.403075	1.151478	0.8708067	1.403075
## 4	-0.1320566	1.403075	1.151478	0.8708067	1.403075
## 5	0.4181071	1.403075	1.151478	0.8708067	1.403075
## 6	-0.3515285	1.403075	1.151478	0.8708067	1.403075

##	venta_Plata	venta_Platino	tamano_empresa
## 1	1.15238	0.8708068	<NA>
## 2	1.15238	0.8708068	<NA>
## 3	1.15238	0.8708068	<NA>
## 4	1.15238	0.8708068	<NA>
## 5	1.15238	0.8708068	<NA>
## 6	1.15238	0.8708068	<NA>

# Capítulo 6. Aplicación de modelo

## 6.1. Modelo regresión Lineal - con todas las variables de compra y venta de oro

En este modelo se intenta explicar los costos y gastos en función de las variables de compra y venta de metales preciosos.

```
library(broom)

mod1 <- lm(costos_gastos_totales ~ compra_Oro + compra_Plata + compra_Platino +
          venta_Oro + venta_Plata + venta_Platino + ingresos_totales,
          data= base_modelo_lineal)
anova(mod1)
```

```
## Analysis of Variance Table
##
## Response: costos_gastos_totales
##           Df Sum Sq Mean Sq F value    Pr(>F)
## compra_Oro      1 0.00121  0.00121   0.0379   0.8462
## compra_Plata     1 0.00359  0.00359   0.1122   0.7388
## ingresos_totales 1 1.64429  1.64429  51.3861 1.098e-09 ***
## Residuals      62 1.98392  0.03200
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

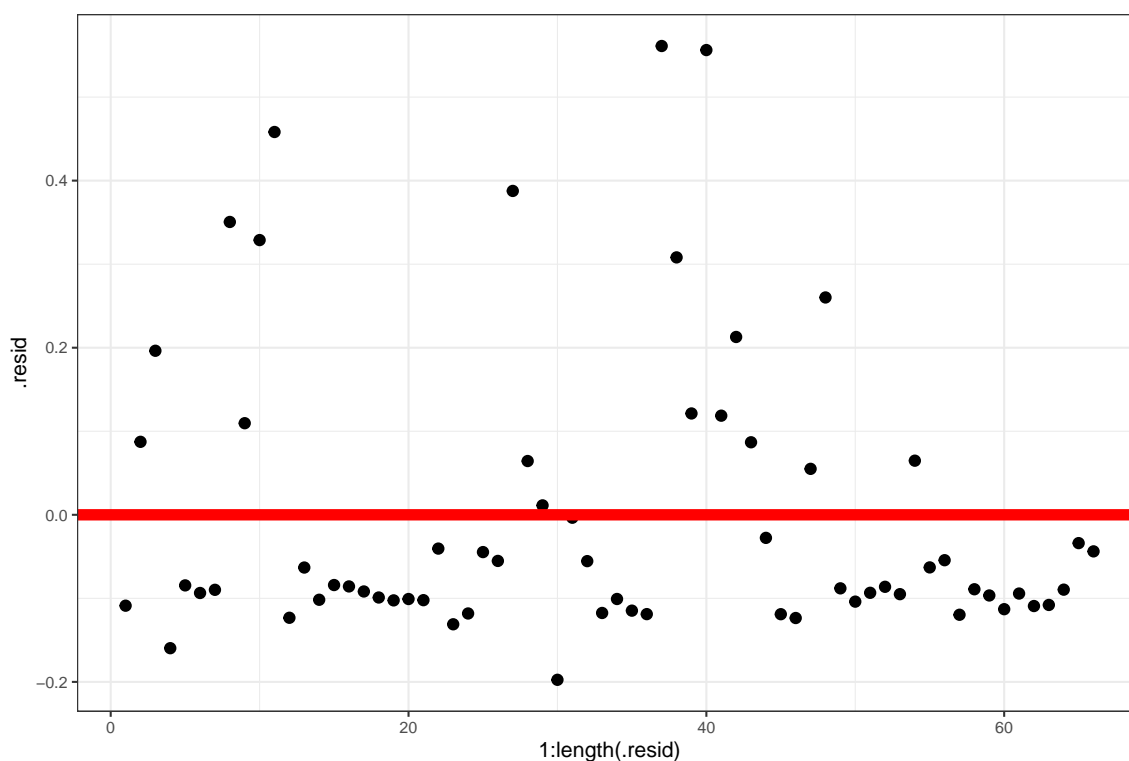
Calculamos el resumen del modelo 1:

```
summary(mod1)

##
## Call:
## lm(formula = costos_gastos_totales ~ compra_Oro + compra_Plata +
##     compra_Platino + venta_Oro + venta_Plata + venta_Platino +
##     ingresos_totales, data = base_modelo_lineal)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.19751 -0.10238 -0.08597  0.06201  0.56114
##
## Coefficients: (4 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.02387    0.03855   0.619   0.538
## compra_Oro     -0.01201    0.03696  -0.325   0.746
## compra_Plata    0.01249    0.03695   0.338   0.736
## compra_Platino      NA          NA      NA      NA
```

```
## venta_Oro          NA          NA          NA          NA
## venta_Plata        NA          NA          NA          NA
## venta_Platino      NA          NA          NA          NA
## ingresos_totales  1.11327    0.15530    7.168    1.1e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1789 on 62 degrees of freedom
## Multiple R-squared:  0.4539, Adjusted R-squared:  0.4275
## F-statistic: 17.18 on 3 and 62 DF,  p-value: 3.124e-08
```

```
a <- augment(mod1)
ggplot(a, aes(x=1:length(.resid), y=.resid))+
  geom_point() +
  geom_hline(yintercept = 0, lwd=2, col= "red")
```



Evaluación de los modelos: AIC, BIC y R2.

El AIC NO es una prueba de hipótesis sobre el ajuste de un modelo, sino más bien un criterio paramétrico comparativo entre modelos y representa por esto una herramienta para selección de modelos. Dado un conjunto de datos, es posible encontrar varios modelos que se ajustan a ellos. La idea es ranquearlos de acuerdo al AIC. El modelo que esté asociado al menor AIC, se considera mejor entre aquellos que se ajustan.

```
glance(mod1)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic p.value    df logLik  AIC   BIC
##   <dbl>      <dbl> <dbl>      <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>
## 1    0.454      0.427 0.179      17.2 3.12e-8     3   22.0 -34.0 -23.1
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

```
library(broom)
library("broom.mixed")

mod1 <- lm(costos_gastos_totales ~ estado, data= base_modelo_lineal)
anova(mod1)
```

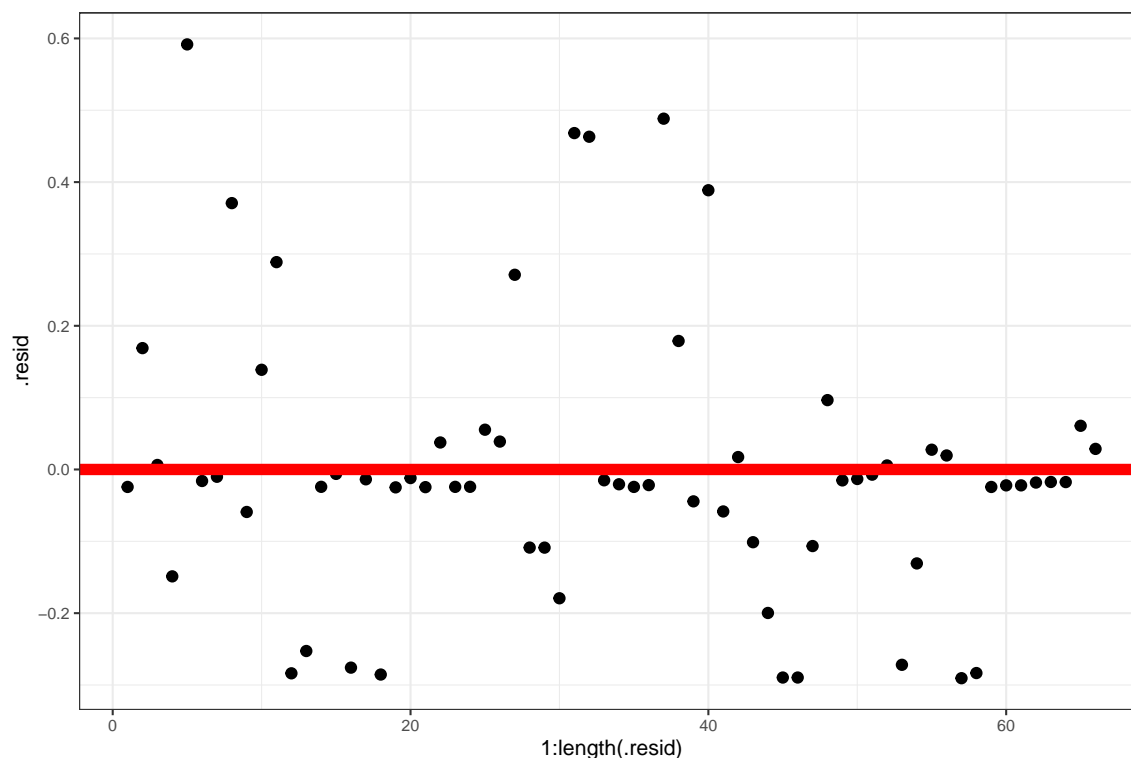
```
## Analysis of Variance Table
##
## Response: costos_gastos_totales
##           Df Sum Sq Mean Sq F value    Pr(>F)
## estado      1  1.1820   1.1820  30.862 5.742e-07 ***
## Residuals  64  2.4511   0.0383
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Calculamos el resumen del modelo 1

```
summary(mod1)
```

```
##
## Call:
## lm(formula = costos_gastos_totales ~ estado, data = base_modelo_lineal)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.29051 -0.09070 -0.01786  0.02839  0.59167
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.33676    0.03407  -9.885 1.66e-14 ***
## estadoVIGILANCIA 0.26764    0.04818   5.555 5.74e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1957 on 64 degrees of freedom
## Multiple R-squared:  0.3253, Adjusted R-squared:  0.3148
## F-statistic: 30.86 on 1 and 64 DF, p-value: 5.742e-07
```

```
a <- augment(mod1)
ggplot(a, aes(x=1:length(.resid), y=.resid))+
  geom_point() +
  geom_hline(yintercept = 0, lwd=2, col= "red")
```



Evaluación de los modelos: AIC, BIC y R2.

```
glance(mod1)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic p.value    df logLik   AIC   BIC
##   <dbl>      <dbl> <dbl>      <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>
## 1    0.325        0.315 0.196      30.9 5.74e-7     1   15.0 -24.0 -17.5
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

## 6.2. Modelo Regresión lineal sin efectos aleatorios

En este modelo utilizamos una fórmula sin el componente aleatorio. Un modelo lineal simple

```
mod2 <- lm(costos_gastos_totales ~ estado + ingresos_totales,
           data= base_modelo_lineal)
anova(mod2)
```

```
## Analysis of Variance Table
##
## Response: costos_gastos_totales
##              Df Sum Sq Mean Sq F value    Pr(>F)
## estado         1 1.18196  1.18196   51.004 1.131e-09 ***
## ingresos_totales 1 0.99111  0.99111   42.769 1.258e-08 ***
## Residuals      63 1.45995  0.02317
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

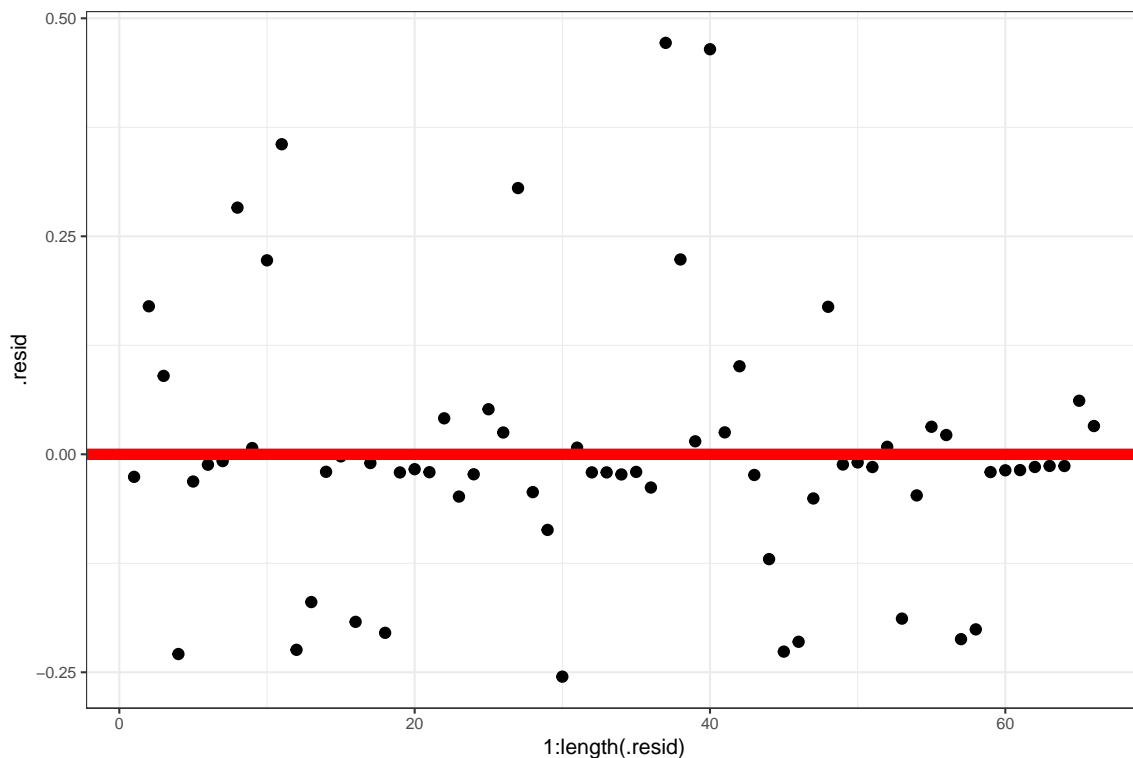
```
summary(mod2)
```

```
##
## Call:
```



```
## lm(formula = costos_gastos_totales ~ estado + ingresos_totales,
##     data = base_modelo_lineal)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.25499 -0.04208 -0.01584  0.02511  0.47185
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.1120     0.0434  -2.580   0.0122 *
## estadoVIGILANCIA  0.1881     0.0394   4.773 1.12e-05 ***
## ingresos_totales  0.9080     0.1389   6.540 1.26e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1522 on 63 degrees of freedom
## Multiple R-squared:  0.5981, Adjusted R-squared:  0.5854
## F-statistic: 46.89 on 2 and 63 DF,  p-value: 3.375e-13
```

```
a <- augment(mod2)
ggplot(a, aes(x=1:length(.resid), y=.resid))+
  geom_point() +
  geom_hline(yintercept = 0, lwd=2, col= "red")
```



Evaluación de los modelos: AIC, BIC y R2.

```
glance(mod2)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic p.value    df logLik   AIC    BIC
##   <dbl>      <dbl> <dbl>      <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl>
##
```

```
## 1      0.598      0.585 0.152      46.9 3.37e-13      2      32.1 -56.2 -47.5
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

### 6.3. Modelo lineal con intercepto aleatorio

En este modelo utilizamos una fórmula con el componente aleatorio por departamento

```
library(lme4)
mod4 <- lmer(costos_gastos_totales ~ ingresos_totales + (1|departamento),
             data= base_modelo_lineal)
```

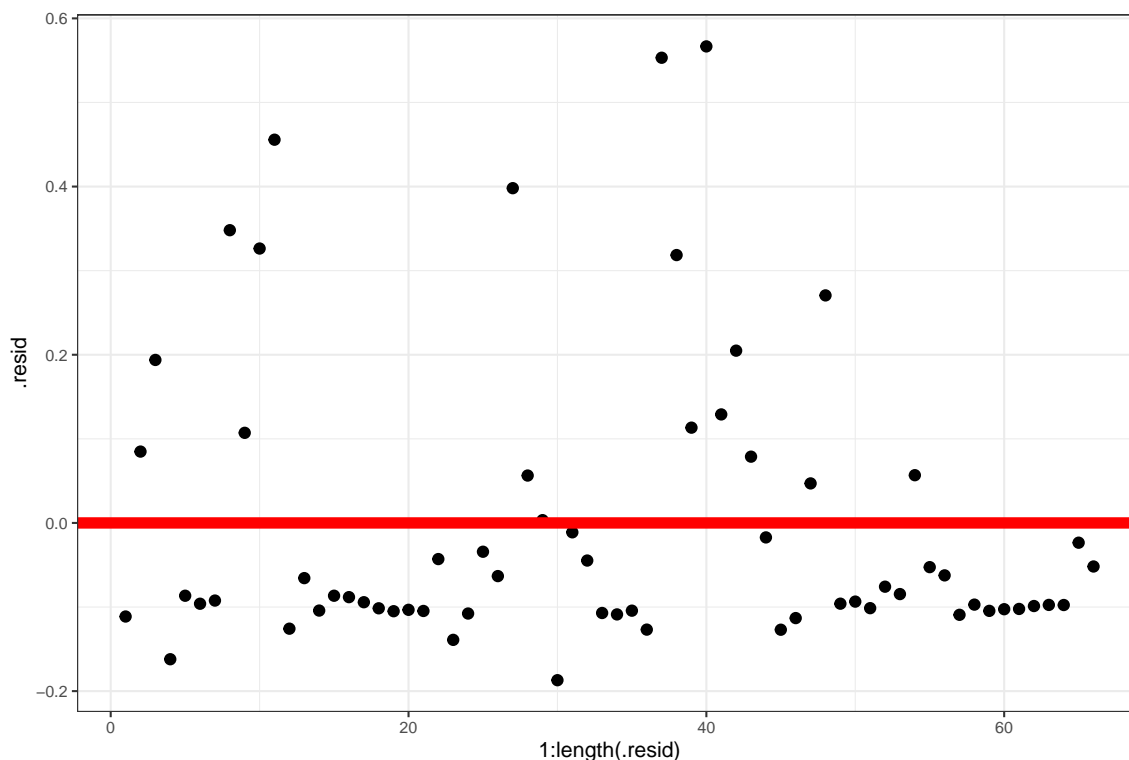
```
anova(mod4)
```

```
## Analysis of Variance Table
##               npar Sum Sq Mean Sq F value
## ingresos_totales      1 1.6452   1.6452   52.967
```

```
summary(mod4)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: costos_gastos_totales ~ ingresos_totales + (1 | departamento)
##      Data: base_modelo_lineal
##
## REML criterion at convergence: -36.1
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.0611 -0.5900 -0.4911  0.3066  3.2155
##
## Random effects:
##      Groups      Name      Variance Std.Dev.
## departamento (Intercept) 6.354e-23 7.971e-12
## Residual              3.106e-02 1.762e-01
## Number of obs: 66, groups:  departamento, 3
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)    0.02375    0.03796    0.626
## ingresos_totales 1.11269    0.15289    7.278
##
## Correlation of Fixed Effects:
##              (Intr)
## ingrss_ttls 0.821
## convergence code: 0
## boundary (singular) fit: see ?isSingular
```

```
a <- broom.mixed::augment(mod4)
ggplot(a, aes(x=1:length(.resid), y=.resid))+
  geom_point() +
  geom_hline(yintercept = 0, lwd=2, col= "red")
```



Evaluación de los modelos: AIC, BIC y R2.

```
broom.mixed::glance(mod4)
```

```
## # A tibble: 1 x 6
##   sigma logLik   AIC   BIC REMLcrit df.residual
##   <dbl>  <dbl> <dbl> <dbl>    <dbl>      <int>
## 1 0.176   18.0 -28.1 -19.3    -36.1         62
```

#### 6.4. Modelo Lineal con intercepto aleatorio a nivel de NIT

En este modelo utilizamos una fórmula con el componente aleatorio por NIT, además contiene una interacción entre los ingresos totales y el estado

```
library(lme4)
mod5 <- lmer(costos_gastos_totales ~ I(ingresos_totales):estado + compra_Oro
            + compra_Plata + compra_Platino + venta_Oro + venta_Plata +
            venta_Platino + ingresos_totales+(1|NIT),
            data= base_modelo_lineal)

anova(mod5)
```

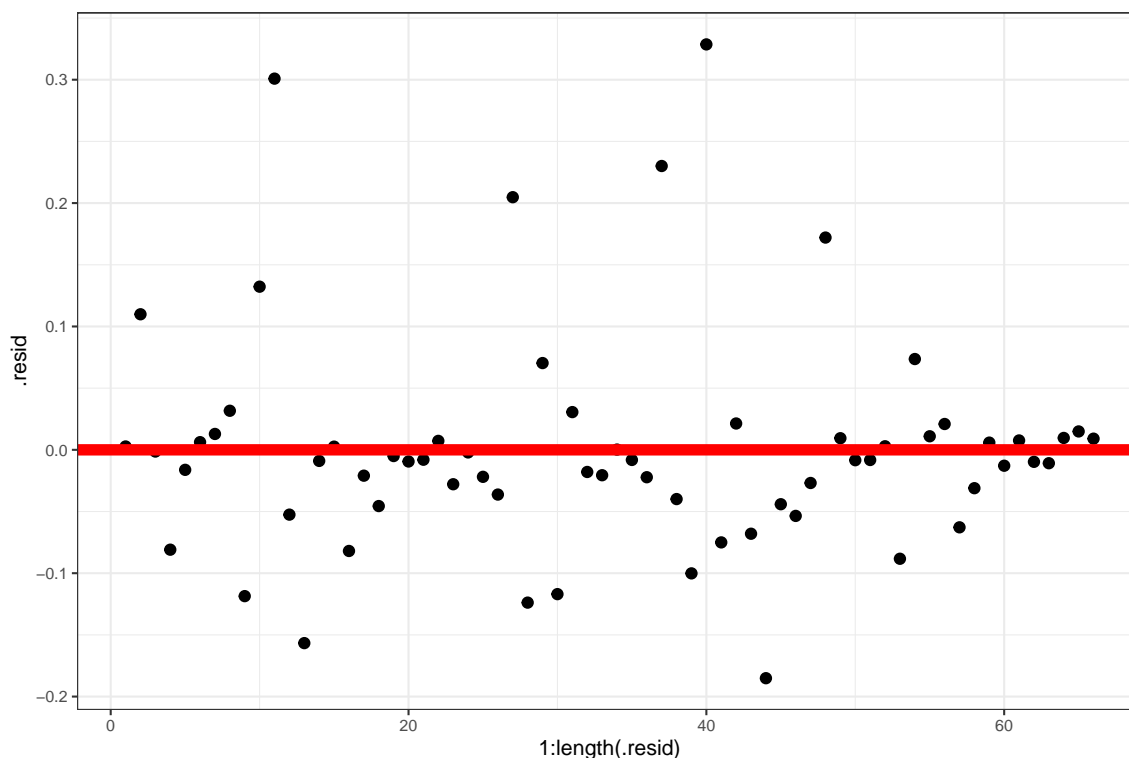
```
## Analysis of Variance Table
##
##               npar  Sum Sq Mean Sq F value
## compra_Oro         1 0.00121 0.00121  0.1053
## compra_Plata        1 0.00359 0.00359  0.3116
## ingresos_totales    1 0.38617 0.38617 33.5135
## I(ingresos_totales):estado 1 0.11916 0.11916 10.3416
```

```
summary(mod5)
```

```
## Linear mixed model fit by REML ['lmerMod']
```

```
## Formula: costos_gastos_totales ~ I(ingresos_totales):estado + compra_Oro +
## compra_Plata + compra_Platino + venta_Oro + venta_Plata +
## venta_Platino + ingresos_totales + (1 | NIT)
## Data: base_modelo_lineal
##
## REML criterion at convergence: -59.1
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.72445 -0.36330 -0.07757  0.08930  3.06148
##
## Random effects:
## Groups      Name      Variance Std.Dev.
## NIT         (Intercept) 0.01317  0.1148
## Residual                0.01152  0.1073
## Number of obs: 66, groups: NIT, 22
##
## Fixed effects:
##
##              Estimate Std. Error t value
## (Intercept)      0.07261    0.05017   1.447
## compra_Oro      -0.01144    0.02220  -0.515
## compra_Plata      0.01410    0.02218   0.636
## ingresos_totales  0.89353    0.19598   4.559
## I(ingresos_totales):estadoINSPECCION 0.75545    0.23492   3.216
##
## Correlation of Fixed Effects:
##              (Intr) cmpr_O cmpr_P ingrs_
## compra_Oro  -0.035
## compra_Plata 0.008 -0.799
## ingrass_ttls 0.622 -0.048 -0.006
## I(:INSPEC 0.341 0.006 0.023 -0.300
## fit warnings:
## fixed-effect model matrix is rank deficient so dropping 5 columns / coefficients
```

```
a <- broom.mixed::augment(mod5)
ggplot(a, aes(x=1:length(.resid), y=.resid))+
  geom_point() +
  geom_hline(yintercept = 0, lwd=2, col= "red")
```



Evaluación de los modelos: AIC, BIC y R2.

```
broom.mixed::glance(mod5)
```

```
## # A tibble: 1 x 6
##   sigma logLik   AIC   BIC REMLcrit df.residual
##   <dbl>   <dbl> <dbl> <dbl>   <dbl>      <int>
## 1 0.107    29.5 -45.1 -29.7   -59.1         59
```

Se verificaron seis modelos. El primero, con una sola variable explicativa llamada “estado” tiene un aporte al costo y al gasto de forma positiva. Este modelo tiene un R2 de 19.85% y un AIC de 74, lo cual nos da el peor modelo. En el segundo modelo, se encuentra que, agregando la variable de ingresos totales y el estado, el modelo presenta un mejor ajuste para la explicación de los costos y gastos (variable objetivo) para el sector minero. Este modelo tiene un R2 de 37.68% y un AIC de 57, con lo cual obtenemos el mejor modelo. El comportamiento de los residuales del modelo 2, también evidencian un mejor comportamiento, al acercarse a cero. Para el tercer modelo, que es el que tiene “departamento” evidenciamos que no es un modelo apropiado para predecir los costos y gastos del sector minero, debido a que las variables no son significativas. Los otros modelos evaluados, no presentan mejoría al ingresarle las variables macroeconómicas estandarizadas y ejecutar el modelo con efectos aleatorios en función de cada compañía y/o el departamento. Por lo anterior, se selecciona como un posible modelo el modelo número 2 sin efecto aleatorio, el cual presenta el menor residual y mejor AIC.



# Capítulo 7. Estimación de esfuerzo

Para las actividades se realiza la siguiente estimación de esfuerzo:

- 1) Consolidación de información: 18h
- 2) Transformación de variables y análisis descriptivo: 10h
- 3) Ajuste y validación de modelos 15h
- 4) Redacción del reporte: 16h





## Capítulo 8. Conclusiones

- Para la variable objetivo propuesta “costos\_gastos\_totales”, se evidencia que no es explicada por las variables macroeconómicas del país, dado que ésta no tiene un efecto volátil durante el año, sino que es constante. Este proceso se ve reflejado en el momento de evaluar las correlaciones con las variables sin transformar, así como con las variables estandarizadas.
- Al adicionar variables económicas relacionadas al sector estudiado, como la compra y venta de materiales preciosos, no se evidencia que éstas expliquen, nuestra variable objetivo planteada.
- El ejercicio de verificación con un modelo de regresión lineal simple, evidencia que ninguna variable, incluyendo las del sector son significativas, obteniendo modelos con  $R^2$  inferiores al 30%, solamente quedando la variable ingresos totales como significativa.
- Se recomienda para los datos trabajados, un acercamiento distinto.
- Cuando los datos están concebidos y correlacionados históricamente, los modelos lineales mixtos son una herramienta muy robusta de análisis estadístico.



# REFERENCIAS

- [https://www.dian.gov.co/ciiu/Documents/Resolucion\\_000139\\_21\\_Nov\\_2012.pdf](https://www.dian.gov.co/ciiu/Documents/Resolucion_000139_21_Nov_2012.pdf)
- <https://linea.ccb.org.co/descripcionciiu/>
- <https://siis.ia.supersociedades.gov.co/>
- [https://www.supersociedades.gov.co/delegatura\\_aec/Paginas/Base-completa-EF-2019.aspx](https://www.supersociedades.gov.co/delegatura_aec/Paginas/Base-completa-EF-2019.aspx)
- [https://www.researchgate.net/publication/314536942\\_Introduccion\\_a\\_los\\_modelos\\_mixtos\\_Introduction\\_to\\_mixed\\_models](https://www.researchgate.net/publication/314536942_Introduccion_a_los_modelos_mixtos_Introduction_to_mixed_models)

