Procesamiento de datos

Marco Andrés Vázquez Hernández

Práctica 1.

Septiembre de 2018

Instituto Politécnico Nacional

Descripción

Instrucciones

1. Integrar los cuatro archivos en una tabla llamada statusVehiculo
2. Trasformar los datos a distancia recorrida y combustible utilizado de cada vehículo por DIA, utilizando los siguientes diagnósticos

a. `'DiagnosticTotalFuelUsedId'`

b. `'DiagnosticOdometerId'`

3. Graficar la relación entre distancia y combustible diario.
4. Sacar el modelo de regresión lineal y = a*x + b
5. Predecir los valores con ruido o sucios con el modelo anterior
6. Contestar lo siguiente:

a. `¿Qué vehículo da mayor rendimiento?`

b. `¿Qué vehículo trabajó mas días?`

c. `¿Cuál es el combustible total utilizado por la flota por semana?`

d. `¿Cuál es la correlación entre ambas variables (distancia, combustible)?`

7. Documentar los resultados

Para la elaboración de la práctica se utilizó la herramienta R Studio.

Integración de los archivos

Se cargaron los datos desde los archivos .csv y se unieron por filas para crear una tabla con las clumnas "vehiculo","fecha","diagnostico" y "valor".

```
wd<-"C:/Users/marco/IPN_BigData/Modulo2/Práctica 1"
setwd(wd)
library(reshape2)
data1<-read.csv("StatusData.csv", header=F)
data2<-read.csv("StatusData2.csv", header=F)
data3<-read.csv("StatusData3.csv", header=F)
data4<-read.csv("StatusData4.csv", header=F)
data<-rbind(data1,data2,data3,data4)
colnames(data)<-c("vehiculo","fecha","diagnostico","valor")
```

Transformación de datos.

Limpiado de datos.

Primero se filtraron los datos para incluir solo las variables "DiagnosticTotalFuelUsedId" y "DiagnosticOdometerId"; se crearon fechas adicionales con los formatos adecuados y se separaron los datos en dos tablas (para cada variable).

```
tidy<-data[data$diagnostico %in% c("DiagnosticTotalFuelUsedId","DiagnosticOdometerId"),]
tidy<-tidy[!duplicated(tidy),]

tidy$fecha2<-gsub("T"," ", tidy$fecha)
```

```
tidy$fecha2<-gsub("Z","", tidy$fecha2)
tidy$fecha2<-as.POSIXct(tidy$fecha2)
tidy$fecha3<-as.Date(tidy$fecha2)

odom<-tidy[tidy$diagnostico=="DiagnosticOdometerId",]
fuel<-tidy[tidy$diagnostico=="DiagnosticTotalFuelUsedId",]
```

Agrupación por día.

Se agruparon los datos por día tomando el valor mínimo del día.

```
aggodom<-aggregate(odom$valor, by=list(odom$vehiculo, odom$fecha3), min)
colnames(aggodom)<-c("vehiculo","fecha3","odom")
aggfuel<-aggregate(fuel$valor, by=list(fuel$vehiculo, fuel$fecha3),min)
colnames(aggfuel)<-c("vehiculo","fecha3","fuel")
```

Normalización de valores por lag diff.

Debido a que los odómetros y la cantidad acumulada de combustible de cada vehículo puede variar de acuerdo a la antiguedad del vehículo, se hizo una normalización por medio de considerar las diferencias entre cada día. En otras palabras, para cada vehículo se ajustó el valor para que en vez de tomar en cuenta el acumulado se tomara la distancia recorrida y lo que consumió de gasolina únicamente en ese día.

Esto hace que el primer día ambos valores; distancia y combustible sean cero. Es decir, debido a que no se tienen los datos del día anterior el primer día se considera el "inicio" y por tanto sus valores serán 0 en ambos casos.

```
# Normalización de valores por lag diff

for (v in unique(aggodom$vehiculo)){
  aux2<-aggodom[aggodom$vehiculo==v,]
  aux2<-aux2[order(aux2$fecha3),]
  aux2$odomdif<-c(0,diff(aux2$odom))
  if(v == "A1"){
    aux3<-aux2
  } else {
    aux3 <- rbind(aux3,aux2)
  }
}
aggodom2<-aux3
aux3<-NULL

for (v in unique(aggfuel$vehiculo)){
  aux2<-aggfuel[aggfuel$vehiculo==v,]
  aux2<-aux2[order(aux2$fecha3),]
  aux2$fueldif<-c(0,diff(aux2$fuel))
  if(v == "A1"){
    aux3<-aux2
  } else {
    aux3 <- rbind(aux3,aux2)
  }
}
```

```
aggfuel2<-aux3
aux3<-NULL
```

Unión de tablas.

Finalmente se unieron ambas tablas por vehículo y fecha y se calcula la variable de rendimiento.

```
dataf<-aggodom2
colnames(dataf)<-c("vehiculo","fecha3","odom","odomdif")
colnames(aggfuel2)<-c("vehiculo","fecha3","fuel","fueldif")
dataf<-merge(dataf, aggfuel2, by=c("vehiculo","fecha3"), all.x=T, all.y=T)

dataf$rendimiento<-dataf$odomdif/dataf$fueldif
```

Por ejemplo, para el vehículo A2 los datos quedan:

```
dataf[dataf$vehiculo=="A2",]
```

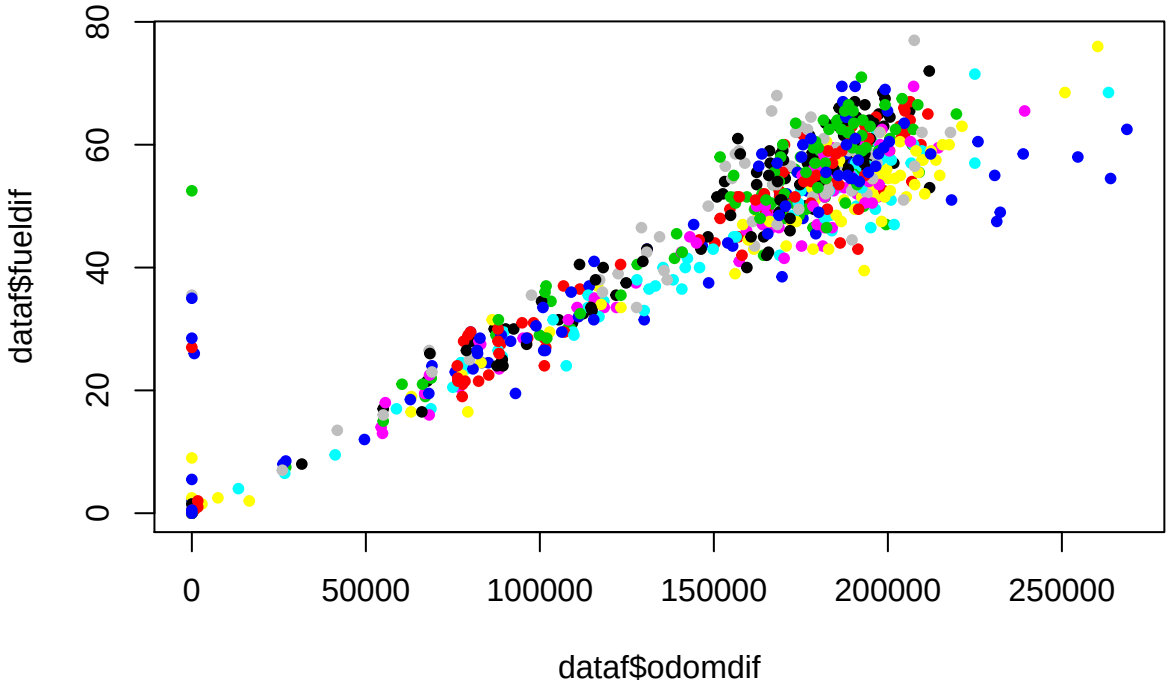
```
##     vehiculo     fecha3      odom odomdif    fuel fueldif
## 429       A2 2018-02-20 110225900       0 38589.5     0.0
## 430       A2 2018-02-21 110334900  109000 38625.5    36.0
## 431       A2 2018-02-22 110449100  114200 38662.5    37.0
## 432       A2 2018-02-23 110637100  188000 38727.0    64.5
## 433       A2 2018-02-24 110827200  190100 38784.0    57.0
## 434       A2 2018-02-25 111002400  175200 38842.0    58.0
## 435       A2 2018-02-26 111103300  100900 38875.5    33.5
## 436       A2 2018-02-27 111202200   98900 38906.0    30.5
## 437       A2 2018-02-28 111406900  204700 38969.5    63.5
## 438       A2 2018-03-01 111569800  162900 39026.0    56.5
## 439       A2 2018-03-02 111769700  199900 39091.5    65.5
## 440       A2 2018-03-03 111968900  199200 39160.5    69.0
## 441       A2 2018-03-04 112157000  188100 39225.5    65.0
## 442       A2 2018-03-05 112237100   80100 39255.0    29.5
## 443       A2 2018-03-06 112428000  190900 39315.0    60.0
## 444       A2 2018-03-07 112596200  168200 39372.0    57.0
## 445       A2 2018-03-08 112790100  193900 39435.0    63.0
## 446       A2 2018-03-09 112977200  187100 39502.0    67.0
## 447       A2 2018-03-10 113165372  188172 39566.5    64.5
## 448       A2 2018-03-11 113263672   98300      NA      NA
## 449       A2 2018-03-12 113263672       0 39601.5    35.0
## 450       A2 2018-03-13 113352572   88900 39630.5    29.0
## 451       A2 2018-03-14 113540572  188000 39689.0    58.5
## 452       A2 2018-03-15 113731172  190600 39758.5    69.5
## 453       A2 2018-03-16 113917972  186800 39828.0    69.5
## 454       A2 2018-03-17 114093572  175600 39888.0    60.0
## 455       A2 2018-03-18 114257372  163800 39946.5    58.5
## 456       A2 2018-03-19 114368472  111100 39978.5    32.0
## 457       A2 2018-03-20 114395472   27000 39987.0     8.5
## 458       A2 2018-03-21 114569472  174000 40042.5    55.5
## 459       A2 2018-03-22 114747372  177900 40103.5    61.0
## 460       A2 2018-03-23 114862972  115600 40144.5    41.0
## 461       A2 2018-03-24 114889172   26200 40152.5     8.0
## 462       A2 2018-03-25 114889172       0      NA      NA
## 463       A2 2018-03-26 114889172       0 40153.0     0.5
```

```
## 464       A2 2018-03-27 114938772   49600 40165.0    12.0
## 465       A2 2018-03-28 115087272  148500 40202.5    37.5
## 466       A2 2018-03-29 115313172  225900 40263.0    60.5
## 467       A2 2018-03-30 115393972   80800 40286.5    23.5
```

Gráfica de relación distancia combustible

La gráfica de la relación distancia-combustible con los diferentes vehículos en colores:

```
plot(dataf$odomdif, dataf$fueldif, pch=20, col=dataf$vehiculo)
```



De donde, tanto en la tabla de ejemplo para el vehículo A1 como en la gráfica se puede observar la presencia de ruido en los datos.

Deteccíon y Tipificación de datos faltantes y Outliers.

Deteccíon y Tipificación de datos faltantes.

Se creó una variable para tipificar las anomalías que se analizaron.

Primero, se tipificaron los datos que aunque sus valores seguramente son atípicos, por definición sabemos que no lo son como aquellos en donde ambos valores (distancia y gasolina) son cero, ya que pueden deberse a que el vehículo se encuentra parado o aquellos casos en donde se toma como inicio en la serie de los datos.

Esto mismo pasa en los casos en donde no hubo cambio en el combustible y por tanto el agrupado de combustible del día se reporta como NA o aquellos en los que simplemente el dato de gasolina no fue registrado en todo el día. Estos casos se tipifican como "Dato faltante gasolina".

También están los casos en donde no se registra cambio en el odómetro pero se registra un cambio en la gasolina; esto puede deberse a que las mediciones de ambas variables no se toman al mismo tiempo y por lo general los casos se presentan, de hecho, después de un "descanso" del vehículo, es decir una anomalía de tipo "Vehículo parado" o una anomalía de tipo "Dato faltante gasolina". Dichos datos se tipifican como "Dato inconsistente distancia".

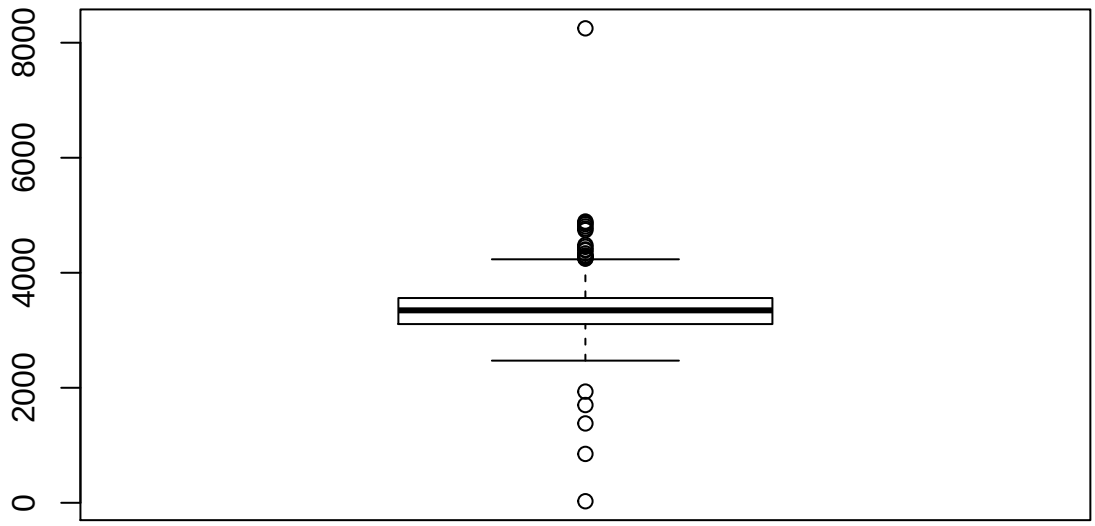
```
# No anomalias
dataf$anomalia<-ifelse((dataf$fueldif==0|is.na(dataf$fueldif))&dataf$odomdif==0,"Vehículo parado/Inicio
                       ifelse(dataf$odomdif==0&dataf$fueldif>0, "Dato inconsistente distancia",
                              ifelse(is.na(dataf$fueldif),"Dato faltante gasolina",
                                     "No anomalía")))
```

Deteccíon y Tipificación de valores anomálicos.

Se analizaron los datos que resultaron no ser faltantes o anomálicos en el sentido antes comentado para verificar si existen valores atípicos los cuales serán descartados del modelaje de predicción.

Se elaboró un boxplot para la variable rendimiento y a partir de ello se determinaron los valores que son atípicos y se clasificaron como "Valor anomálico".

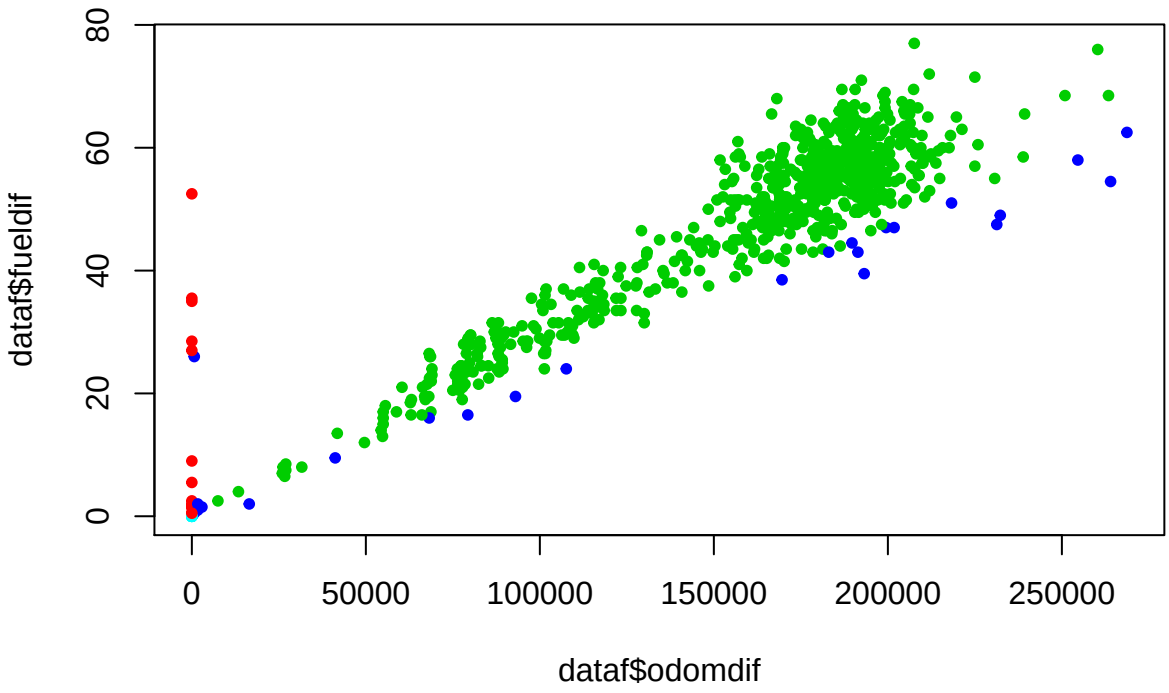
```
# Valores anomálicos

plo<-dataf[dataf$anomalia=="No anomalía",]
bxo<-boxplot(plo$rendimiento)
```




```
dataf$anomalia<-ifelse(dataf$rendimiento %in% bxo$out[bxo$out!=0], "Valor anomálico",dataf$anomalia)
dataf$anomalia<-as.factor(dataf$anomalia)
```

La gráfica con los datos anomálicos en rojo (Datos faltantes) y azul (valores anomálicos) queda:

```
plot(dataf$odomdif, dataf$fueldif, col=dataf$anomalia, pch=20)
```



Modelo de regresión

Se tomaron todos los datos que no fueran anomalías por datos faltantes o valores atípicos y se elaboró un modelo de regresión lineal forzando el intercepto en 0,0 ya que es lógico que con 0 gasolina se recorren 0 unidades de distancia y viceversa análogamente. Se obtienen los siguientes resultados:

```
tidy2<-dataf[dataf$anomalia=="No anomalía",]

lmod <- lm(tidy2$fueldif ~ 0+ tidy2$odomdif)
summary(lmod)
```

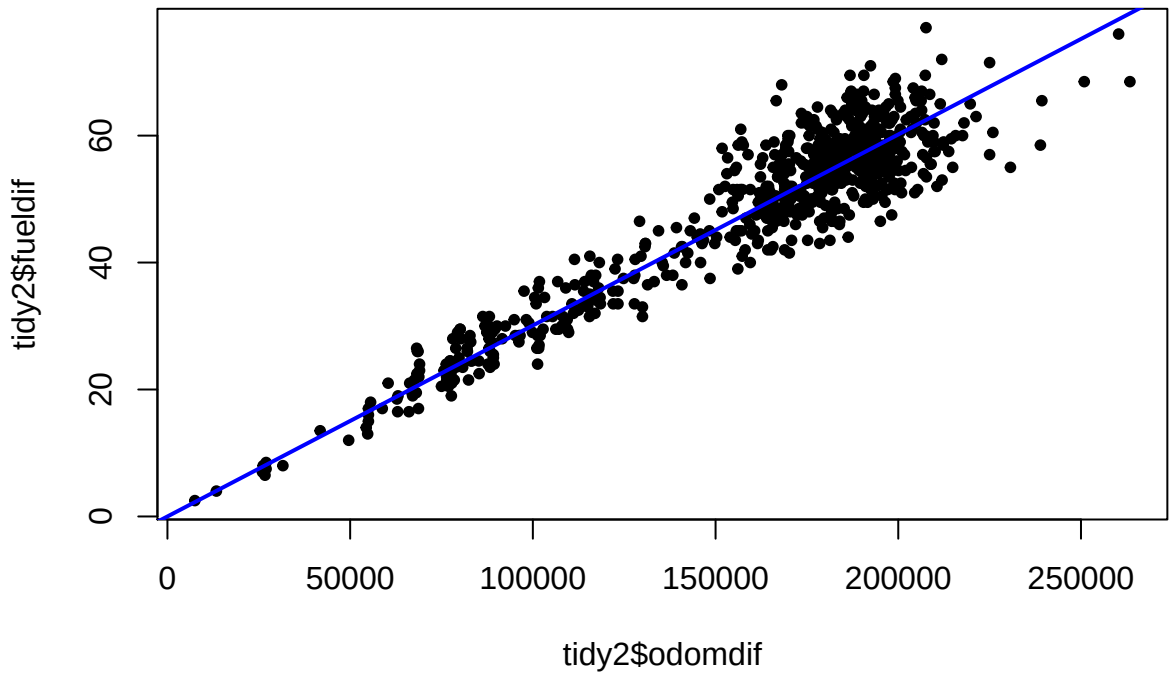
```
##
## Call:
## lm(formula = tidy2$fueldif ~ 0 + tidy2$odomdif)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.4144  -2.8814  -0.3086   3.1749  17.4211
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## tidy2$odomdif 3.009e-04  1.094e-06   275.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.815 on 695 degrees of freedom
```

```
## Multiple R-squared:  0.9909, Adjusted R-squared:  0.9909
## F-statistic: 7.566e+04 on 1 and 695 DF,  p-value: < 2.2e-16
```

Donde se puede observar que el p-value es muy cercano a 0 y el valor de la prueba R cuadrada y Rcuadrada ajustada es muy cercano a 1. Nota: Las demás pruebas estadísticas como lo son la prueba de normalidad de los residuos, entre otras, salen del alcance de esta práctica.

La gráfica del modelo queda:

```
plot(tidy2$odomdif, tidy2$fueldif, pch=20)
abline(lmod, lwd=2, col="blue")
```



Imputación de valores

Si la tabla fuera a ser usada para alguna apllicación y/o reporte diario de dichas métricas de la flota sería recomendable imputar los valores faltantes:

1. En los casos en los que el dato del gasolina falta o en los casos en que el vehículo estuvo parado, se puede suponer que las mediciones no coincidieron y por tanto es mejor estimar de acuerdo a la regresión lineal antes hecha.
2. En los casos en los que el dato de la distancia es cero pero la gasolina no, también se puede asumir que los datos no coincidieron.
3. Se calcula de nuevo el rendimiento.
4. En el caso de los valores atípicos no se recomienda imputar los valores ya que es muy probable que se deba a que el vehículo estuvo en tráfico e imputar esos valores sería perder información real.

```
# Imputación de valores

supertidy<-dataf[,-c(3,5)]
supertidy$fueldif<-ifelse(supertidy$anomalia %in% c("Vehículo parado/Inicio","Dato faltante gasolina"),
                      lmod$coefficients[[1]]*supertidy$odomdif,
                      supertidy$fueldif)
supertidy$odomdif<-ifelse(supertidy$anomalia %in% c("Dato inconsistente distancia"),
                          supertidy$fueldif/lmod$coefficients[[1]],
                          supertidy$odomdif)
supertidy$rendimiento<-supertidy$odomdif/supertidy$fueldif
```

En el ejemplo anterior del vehículo A2 la tabla "limpia" con datos imputados queda:

```
supertidy[supertidy$vehiculo=="A2",]
```

```
##     vehiculo     fecha3   odomdif  fueldif rendimiento
## 429       A2 2018-02-20      0.00  0.00000         NaN
## 430       A2 2018-02-21 109000.00 36.00000    3027.778
## 431       A2 2018-02-22 114200.00 37.00000    3086.486
## 432       A2 2018-02-23 188000.00 64.50000    2914.729
## 433       A2 2018-02-24 190100.00 57.00000    3335.088
## 434       A2 2018-02-25 175200.00 58.00000    3020.690
## 435       A2 2018-02-26 100900.00 33.50000    3011.940
## 436       A2 2018-02-27  98900.00 30.50000    3242.623
## 437       A2 2018-02-28 204700.00 63.50000    3223.622
## 438       A2 2018-03-01 162900.00 56.50000    2883.186
## 439       A2 2018-03-02 199900.00 65.50000    3051.908
## 440       A2 2018-03-03 199200.00 69.00000    2886.957
## 441       A2 2018-03-04 188100.00 65.00000    2893.846
## 442       A2 2018-03-05  80100.00 29.50000    2715.254
## 443       A2 2018-03-06 190900.00 60.00000    3181.667
## 444       A2 2018-03-07 168200.00 57.00000    2950.877
## 445       A2 2018-03-08 193900.00 63.00000    3077.778
## 446       A2 2018-03-09 187100.00 67.00000    2792.537
## 447       A2 2018-03-10 188172.00 64.50000    2917.395
## 448       A2 2018-03-11  98300.00 29.57707    3323.520
## 449       A2 2018-03-12 116323.20 35.00000    3323.520
## 450       A2 2018-03-13  88900.00 29.00000    3065.517
## 451       A2 2018-03-14 188000.00 58.50000    3213.675
## 452       A2 2018-03-15 190600.00 69.50000    2742.446
## 453       A2 2018-03-16 186800.00 69.50000    2687.770
## 454       A2 2018-03-17 175600.00 60.00000    2926.667
## 455       A2 2018-03-18 163800.00 58.50000    2800.000
## 456       A2 2018-03-19 111100.00 32.00000    3471.875
## 457       A2 2018-03-20  27000.00  8.50000    3176.471
## 458       A2 2018-03-21 174000.00 55.50000    3135.135
## 459       A2 2018-03-22 177900.00 61.00000    2916.393
## 460       A2 2018-03-23 115600.00 41.00000    2819.512
## 461       A2 2018-03-24  26200.00  8.00000    3275.000
## 462       A2 2018-03-25      0.00  0.00000         NaN
## 463       A2 2018-03-26   1661.76  0.50000    3323.520
## 464       A2 2018-03-27  49600.00 12.00000    4133.333
## 465       A2 2018-03-28 148500.00 37.50000    3960.000
## 466       A2 2018-03-29 225900.00 60.50000    3733.884
```

```
## 467       A2 2018-03-30  80800.00 23.50000    3438.298
##                          anomalia
## 429       Vehículo parado/Inicio
## 430                   No anomalía
## 431                   No anomalía
## 432                   No anomalía
## 433                   No anomalía
## 434                   No anomalía
## 435                   No anomalía
## 436                   No anomalía
## 437                   No anomalía
## 438                   No anomalía
## 439                   No anomalía
## 440                   No anomalía
## 441                   No anomalía
## 442                   No anomalía
## 443                   No anomalía
## 444                   No anomalía
## 445                   No anomalía
## 446                   No anomalía
## 447                   No anomalía
## 448       Dato faltante gasolina
## 449 Dato inconsistente distancia
## 450                   No anomalía
## 451                   No anomalía
## 452                   No anomalía
## 453                   No anomalía
## 454                   No anomalía
## 455                   No anomalía
## 456                   No anomalía
## 457                   No anomalía
## 458                   No anomalía
## 459                   No anomalía
## 460                   No anomalía
## 461                   No anomalía
## 462       Vehículo parado/Inicio
## 463 Dato inconsistente distancia
## 464                   No anomalía
## 465                   No anomalía
## 466                   No anomalía
## 467                   No anomalía
```

Preguntas.

Nota: Se tomaron los datos limpios, es decir sin considerar anomalías excepto en el inciso c.

a. ¿Qué vehículo da mayor rendimiento?

```
# vehículo con mayor rendimiento
aux<-aggregate(plo$rendimiento, by=list(plo$vehiculo), mean)
colnames(aux)<-c("Vehículo","rendimiento_promedio")
aux<-aux[order(aux$rendimiento_promedio, decreasing=T),]
aux
```

```
##    Vehículo rendimiento_promedio
```

```
## 15       A3             3709.865
## 7       A15             3630.158
## 6       A14             3609.469
## 4       A12             3606.524
## 13      A20             3588.235
## 5       A13             3558.640
## 10      A18             3479.436
## 17       A6             3441.529
## 20       A9             3423.880
## 3       A11             3405.619
## 16       A5             3390.432
## 14      A21             3360.189
## 2       A10             3263.025
## 11      A19             3198.678
## 19       A8             3146.800
## 12       A2             3109.128
## 18       A7             3107.599
## 9       A17             3096.768
## 1        A1             3076.492
## 8       A16             3012.026
```

b. ¿Qué vehículo trabajó mas días?

```
# Vehículo que trabajó más días
aux<-aggregate(plo$fecha3, by=list(plo$vehiculo), min)
colnames(aux)<-c("Vehículo","min")
aux2<-aggregate(plo$fecha3, by=list(plo$vehiculo), max)
colnames(aux2)<-c("Vehículo","max")
aux3<-merge(aux, aux2, by="Vehículo")
aux3$dif<-aux3$max-aux3$min
aux3<-aux3[order(aux3$dif, decreasing=T),]
aux3
```

```
##    Vehículo        min        max     dif
## 7       A15 2018-02-21 2018-03-31 38 days
## 11      A19 2018-02-21 2018-03-31 38 days
## 2       A10 2018-02-22 2018-03-31 37 days
## 3       A11 2018-02-22 2018-03-31 37 days
## 4       A12 2018-02-22 2018-03-31 37 days
## 5       A13 2018-02-22 2018-03-31 37 days
## 6       A14 2018-02-22 2018-03-31 37 days
## 8       A16 2018-02-22 2018-03-31 37 days
## 9       A17 2018-02-21 2018-03-30 37 days
## 12       A2 2018-02-21 2018-03-30 37 days
## 13      A20 2018-02-21 2018-03-30 37 days
## 14      A21 2018-02-22 2018-03-31 37 days
## 15       A3 2018-02-22 2018-03-31 37 days
## 16       A5 2018-02-22 2018-03-31 37 days
## 17       A6 2018-02-21 2018-03-30 37 days
## 18       A7 2018-02-22 2018-03-31 37 days
## 19       A8 2018-02-22 2018-03-30 36 days
## 20       A9 2018-02-22 2018-03-30 36 days
## 1        A1 2018-02-21 2018-03-28 35 days
## 10      A18 2018-02-22 2018-03-29 35 days
```

c. ¿Cuál es el combustible total utilizado por la flota por semana?

```r
# Combustible total de la flota por semana

library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```r
aux<-fuel
aux$semana<-week(ymd(as.character(aux$fecha3)))

totalv<-aggregate(aux$valor, by=list(aux$vehiculo, aux$semana), min)
colnames(totalv)<-c("v","s","min")
aux2<-aggregate(aux$valor, by=list(aux$vehiculo, aux$semana), max)
colnames(aux2)<-c("v","s","max")
totalv<-merge(totalv, aux2, by=c("v","s"))
totalv$dif<-totalv$max-totalv$min

totalf<-aggregate(totalv$dif, by=list(totalv$s), sum)
colnames(totalf)<-c("Semana","Total_gasolina")
totalf
```

```
##   Semana Total_gasolina
## 1      8         4651.5
## 2      9         6647.0
## 3     10         6589.0
## 4     11         6586.5
## 5     12         5964.0
## 6     13         4315.0
```

d. ¿Cuál es la correlación entre ambas variables (distancia, combustible)? Correlación de Pearson:

```r
cor(plo$odomdif, plo$fueldif)
```

```
## [1] 0.9278223
```