Benemérita Universidad Autónoma de Puebla
Facultad de Ciencias de la Computación

# Máquinas de aprendizaje

## Reporte: Tarea de regresión


**Docente: Abraham Sánchez López**



| Alumno | Matrícula |
|---|---|
| Taisen Romero Bañuelos | 202055209 |

# Proyecto 1 - Regresión

## Predicción de tarifas aéreas en nuevas rutas

Empecemos con el análisis exploratorio de los datos (el general, no el específico que se solicita en el inciso A).

```
> head(data)
  S_CODE               S_CITY E_CODE                E_CITY COUPON NEW VACATION  SW      HI S_INCOME E_INCOME   S_POP   E_POP
1      * Dallas/Fort Worth    TX      * Amarillo            TX   1.00   3       No Yes 5291.99    28637    21112 3036732  205711
2      * Atlanta              GA      * Baltimore/Wash Intl MD   1.06   3       No  No 5419.16    26993    29838 3532657 7145897
3      * Boston               MA      * Baltimore/Wash Intl MD   1.06   3       No  No 9185.28    30124    29838 5787293 7145897
4    ORD Chicago              IL      * Baltimore/Wash Intl MD   1.06   3       No Yes 2657.35    29260    29838 7830332 7145897
5    MDW Chicago              IL      * Baltimore/Wash Intl MD   1.06   3       No Yes 2657.35    29260    29838 7830332 7145897
6      * Cleveland            OH      * Baltimore/Wash Intl MD   1.01   3       No Yes 3408.11    26046    29838 2230955 7145897
        SLOT GATE DISTANCE   PAX   FARE
1       Free Free      312  7864  64.11
2       Free Free      576  8820 174.47
3       Free Free      364  6452 207.76
4 Controlled Free      612 25144  85.47
5       Free Free      612 25144  85.47
6       Free Free      309 13386  56.76
> sum(is.na(data))
[1] 0
```

Al parecer no hay datos nulos, pero en E_CODE y S_CODE hay muchas observaciones con asteriscos en lugar de un código como tal. También, S_CITY tiene el nombre de la ciudad de origen y lo que parece ser una abreviación del estado (de origen, también).

```
> str(data)
'data.frame':   638 obs. of  18 variables:
 $ S_CODE   : chr  "*" "*" "*" "ORD" ...
 $ S_CITY   : chr  "Dallas/Fort Worth   TX" "Atlanta             GA" "Boston              MA" "Chicago             IL" ...
 $ E_CODE   : chr  "*" "*" "*" "*" ...
 $ E_CITY   : chr  "Amarillo            TX" "Baltimore/Wash Intl MD" "Baltimore/Wash Intl MD" "Baltimore/Wash Intl MD" ...
 $ COUPON   : num  1 1.06 1.06 1.06 1.06 1.01 1.28 1.15 1.33 1.6 ...
 $ NEW      : int  3 3 3 3 3 3 3 3 3 2 ...
 $ VACATION: chr  "No" "No" "No" "No" ...
 $ SW       : chr  "Yes" "No" "No" "Yes" ...
 $ HI       : num  5292 5419 9185 2657 2657 ...
 $ S_INCOME: num  28637 26993 30124 29260 29260 ...
 $ E_INCOME: num  21112 29838 29838 29838 29838 ...
 $ S_POP    : int  3036732 3532657 5787293 7830332 7830332 2230955 3036732 1440377 3770125 1694803 ...
 $ E_POP    : int  205711 7145897 7145897 7145897 7145897 7145897 7145897 7145897 7145897 7145897 ...
 $ SLOT     : chr  "Free" "Free" "Free" "Controlled" ...
 $ GATE     : chr  "Free" "Free" "Free" "Free" ...
 $ DISTANCE: int  312 576 364 612 612 309 1220 921 1249 964 ...
 $ PAX      : int  7864 8820 6452 25144 25144 13386 4625 5512 7811 4657 ...
 $ FARE     : num  64.1 174.5 207.8 85.5 85.5 ...
>
```

La mayoría de las variables de tipo char las podríamos convertir en factores para hacer más fácil el trabajo.

```
> summary(data)
    S_CODE             S_CITY             E_CODE             E_CITY             COUPON           NEW          VACATION
 Length:638         Length:638         Length:638         Length:638         Min.   :1.000   Min.   :0.000   Length:638
 Class :character   Class :character   Class :character   Class :character   1st Qu.:1.040   1st Qu.:3.000   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :1.150   Median :3.000   Mode  :character
                                                                             Mean   :1.202   Mean   :2.754
                                                                             3rd Qu.:1.298   3rd Qu.:3.000
                                                                             Max.   :1.940   Max.   :3.000
     SW                  HI            S_INCOME        E_INCOME         S_POP             E_POP             SLOT
 Length:638         Min.   : 1230   Min.   :14600   Min.   :14600   Min.   :  29838   Min.   : 111745   Length:638
 Class :character   1st Qu.: 3090   1st Qu.:24706   1st Qu.:23903   1st Qu.:1862106   1st Qu.:1228816   Class :character
 Mode  :character   Median : 4208   Median :28637   Median :26409   Median :3532657   Median :2195215   Mode  :character
                    Mean   : 4442   Mean   :27760   Mean   :27664   Mean   :4557004   Mean   :3194503
                    3rd Qu.: 5481   3rd Qu.:29694   3rd Qu.:31981   3rd Qu.:7830332   3rd Qu.:4549784
                    Max.   :10000   Max.   :38813   Max.   :38813   Max.   :9056076   Max.   :9056076
    GATE              DISTANCE          PAX             FARE
 Length:638         Min.   : 114.0   Min.   : 1504   Min.   : 42.47
 Class :character   1st Qu.: 455.0   1st Qu.: 5328   1st Qu.:106.29
 Mode  :character   Median : 850.0   Median : 7792   Median :144.60
                    Mean   : 975.7   Mean   :12782   Mean   :160.88
                    3rd Qu.:1306.2   3rd Qu.:14090   3rd Qu.:209.35
                    Max.   :2764.0   Max.   :73892   Max.   :402.02
>
```

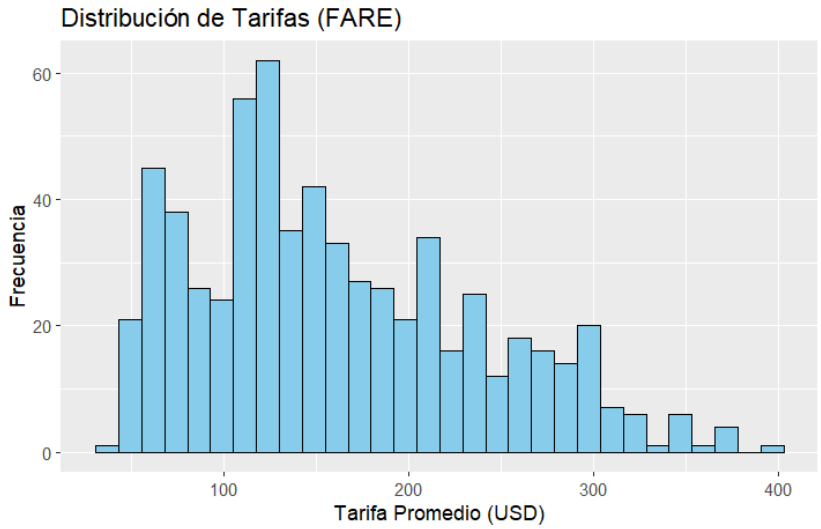
Aquí podemos notar algunas cosas interesantes, por ejemplo, si comparamos el primer y tercer cuartil de las variables de ingresos personales de la ciudad de origen vs la de destino (S_INCOME vs E_INCOME) podemos notar que hay una tendencia de viajar a lugares de mayores ingresos a los de la ciudad de origen, esto como un efecto migratorio pero para nuestro problema podría ser de interés observar su correlación con las tarifas. Y hablando de tarifas, la tarifa promedio es de $160.88, lo cual es un precio accesible considerando el mínimo y el máximo que hay. Las tarifas cercanas al mínimo de FARE ($42.47) podrían ser vuelos cortos o con mucha competencia (para poder competir tienen que bajar el precio), y por el otro lado, las tarifas cercanas al máximo ($402.02) pueden ser tarifas de viajes largos o de rutas monopolizadas (poca competencia. Observar variable HI) o con poca oferta (observar variable PAX), aunque también podría aumentar el precio por los congestionamientos y la escasez de SLOTS (horarios).

A ciencia cierta no sabemos qué representan los asteriscos en las variables S_CODE y E_CODE, yo puedo suponer que un asterisco significa que la ciudad no tiene un código, o que se trata de un aeropuerto desconocido o no especificado, por lo que considero adecuado reemplazar los "*" por un "UNK" de desconocido, ya que podría tratarse de un código desconocido o genérico (o simplemente que el aeropuerto no tenga código, pero como no sabemos lo clasificamos como desconocido). Este enfoque permite mantener la integridad de los datos, pues, si simplemente elimino los asteriscos podría eliminar información útil. Pero para ver si hay que tratar más la información hay que revisar la proporción de códigos desconocidos, ya que muchos valores UNK podrían afectar al modelo.

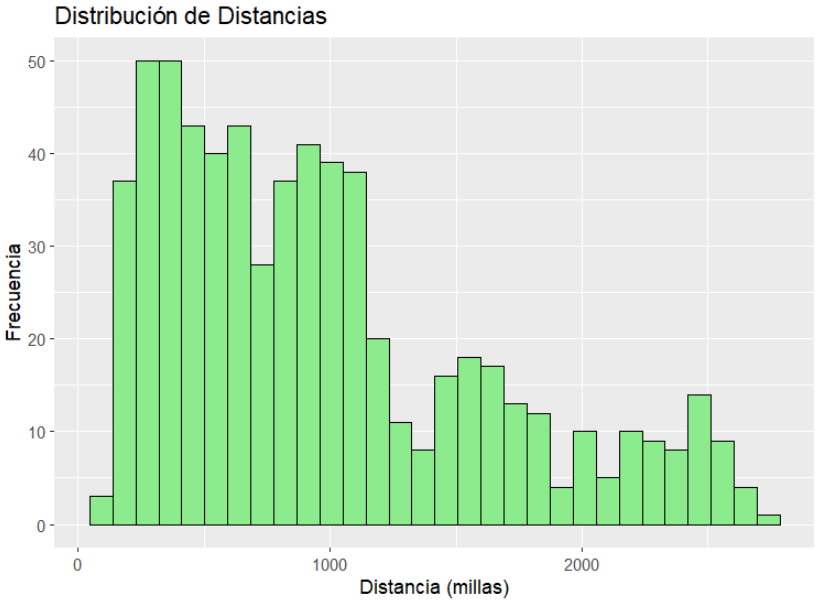
```
> #Contar valores "UNK"
> table(data$S_CODE == "UNK")

FALSE  TRUE
  184   454
> table(data$E_CODE == "UNK")

FALSE  TRUE
  137   501
>
```

La mayoría de las observaciones tienen la etiqueta UNK. Esto podría ser problemático si el código de origen y destino del aeropuerto fuera de relevancia. Afortunadamente las variables de ciudad de origen y destino están completas, por lo que podríamos darnos el lujo de descartar los códigos del aeropuerto, pues, gracias a las variables S_CITY, E_CITY tenemos una referencia del origen y destino de los vuelos. Si fuera necesario el código del aeropuerto simplemente podría crear un data frame que incluya los códigos.

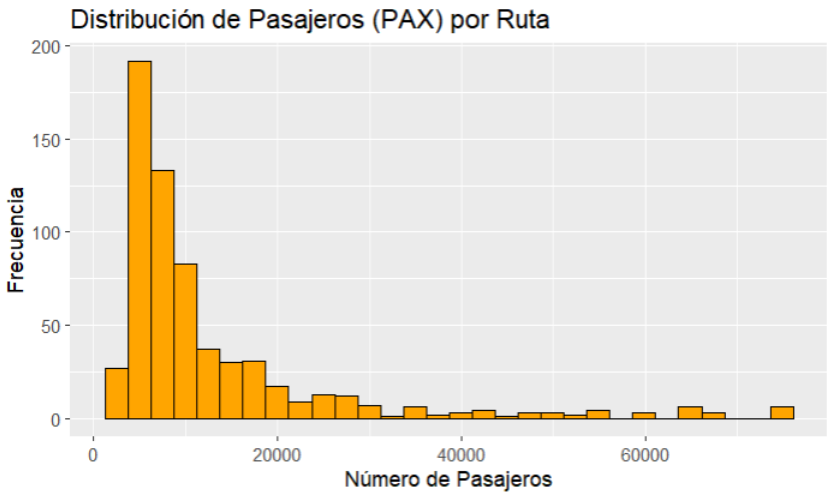
En fin, luego de convertir en factor las variables categóricas que eran de tipo char me di a la tarea de checar la distribución de algunas variables numéricas.



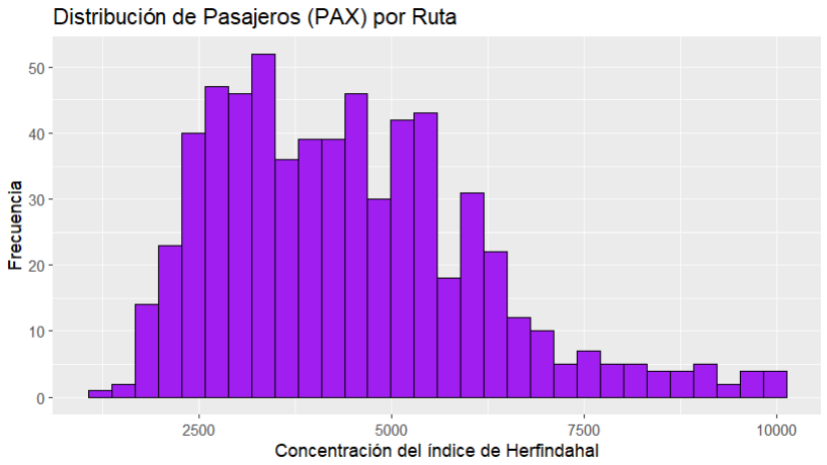

Esta imagen corrobora lo que notamos con la función summary(), la mayoría de las tarifas se cobran en $160.




Aquí vemos que la mayoría de viajes tienen un recorrido de menos de 1,300 millas. Si lo relacionamos con el histograma anterior podemos entender que la mayoría de los viajes son relativamente económicos porque no recorren una distancia considerable.




Y si ahora combinamos lo que sabemos de los tres histogramas combinado con el análisis de la función summary() podemos reforzar la idea de que la mayoría de la gente que viaja lo hace en busca de mejores oportunidades, pues recordemos que la mayoría de los viajes tenían como destino ciudades con mayores ingresos económicos. A su vez, podríamos reforzar la idea de que en promedio, los precios son accesibles debido a la alta demanda (lo que posiblemente también genera competencia).

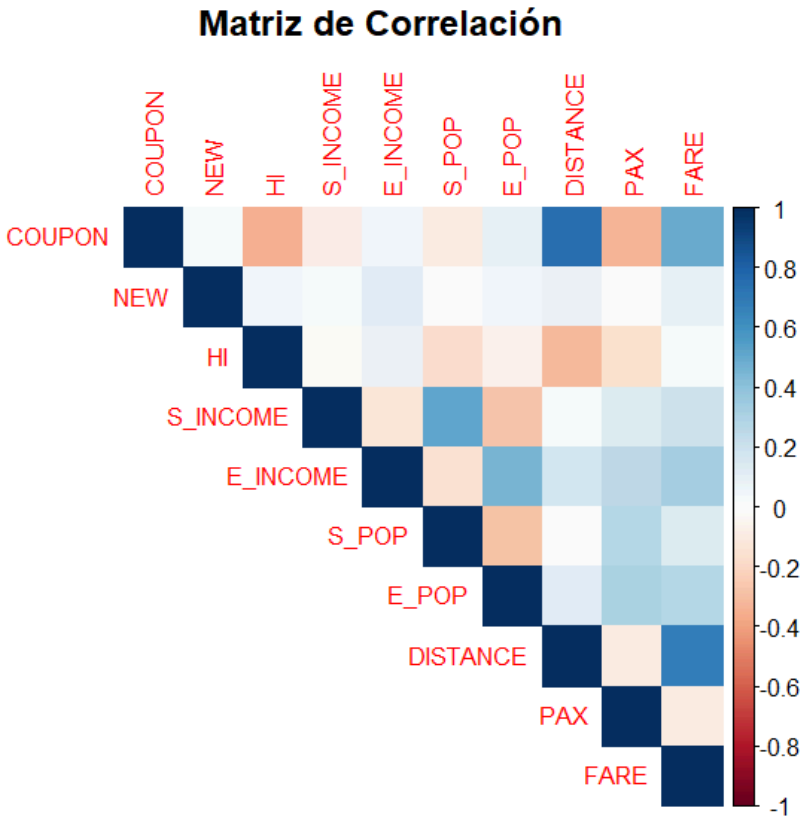



Pues en efecto, hay una alta competitividad para la mayoría de los vuelos tomados (un valor cercano al máximo 10,000 sugiere que hay poca competencia, y un valor menor de HI sugiere más competencia).

Ahora revisemos la correlación entre variables.

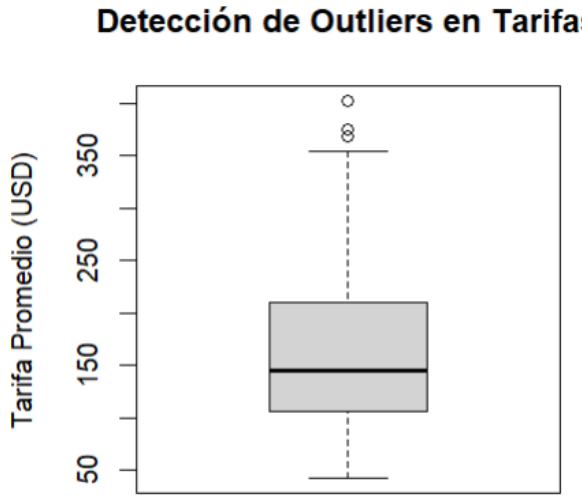
```
> cor_matrix
              COUPON         NEW          HI    S_INCOME    E_INCOME        S_POP        E_POP    DISTANCE         PAX        FARE
COUPON     1.00000000  0.02022307 -0.34725207 -0.08840265  0.0468892 -0.10776336  0.09496994  0.74680521 -0.33697358  0.49653696
NEW        0.02022307  1.00000000  0.05414685  0.02659673  0.1133766 -0.01667212  0.05856818  0.08096520  0.01049527  0.09172969
HI        -0.34725207  0.05414685  1.00000000 -0.02738221  0.0823926 -0.17249541 -0.06245600 -0.31237457 -0.16896078  0.02519492
S_INCOME  -0.08840265  0.02659673 -0.02738221  1.00000000 -0.1388642  0.51718718 -0.27228027  0.02815334  0.13819710  0.20913485
E_INCOME   0.04688920  0.11337664  0.08239260 -0.13886420  1.0000000 -0.14405857  0.45841806  0.17653074  0.25996105  0.32609229
S_POP     -0.10776336 -0.01667212 -0.17249541  0.51718718 -0.1440586  1.00000000 -0.28014283  0.01843667  0.28461056  0.14509708
E_POP      0.09496994  0.05856818 -0.06245600 -0.27228027  0.4584181 -0.28014283  1.00000000  0.11563970  0.31469750  0.28504299
DISTANCE   0.74680521  0.08096520 -0.31237457  0.02815334  0.1765307  0.01843667  0.11563970  1.00000000 -0.10248160  0.67001599
PAX       -0.33697358  0.01049527 -0.16896078  0.13819710  0.2599611  0.28461056  0.31469750 -0.10248160  1.00000000 -0.09070541
FARE       0.49653696  0.09172969  0.02519492  0.20913485  0.3260923  0.14509708  0.28504299  0.67001599 -0.09070541  1.00000000
> 
```

Aquí podemos corroborar que la distancia está correlacionada con la competitividad y los cupones. También el número de pasajes tiene que ver con los cupones, lo que me hace pensar que para competir lo que hacen es ofrecer ofertas para aquellos viajes donde hay mucha gente. Lo más importante, creo yo, es que la tarifa tiene una mayor correlación con la distancia, los cupones y la ciudad de destino.

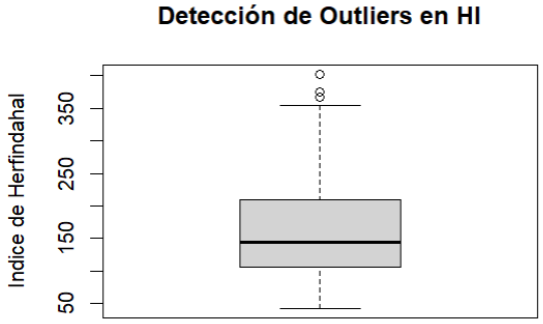


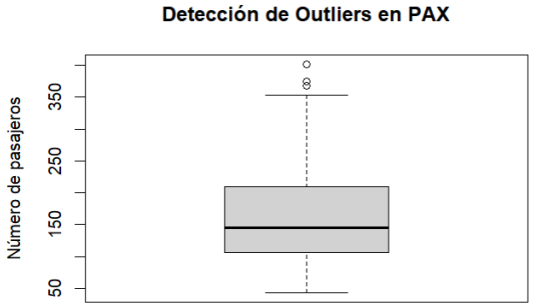

Realizando diagramas de caja para algunas de las variables noté que habían datos atípicos, por lo que habrá que hacer algo al respecto ya que pueden afectar nuestros modelos de regresión. Por ejemplo, en la regresión lineal los datos atípicos (outliers) pueden inclinar la pendiente ya que como se usan errores al cuadrado, los puntos alejados tienen un impacto más grande en el modelo. Y en la regresión logística aunque aguanta mejor los outliers para las variables dependientes (a menos que sean muchos los datos atípicos), todavía tiene complicaciones si hay outliers en las variables independientes, pues, puede empujar las probabilidades hacia 1 o 0 porque aumenta la varianza. Por ejemplo, si una ruta tiene una tarifa extremadamente alta en

presencia de Southwest, el modelo puede asumir incorrectamente que la presencia de Southwest aumenta las tarifas, cuando normalmente debería bajarlas. Capaz que ese outlier existe porque hubo inusuales restricciones (SLOT y GATE) y eso aumentó el precio debido a la poca oferta y alta demanda, pero no es más que una situación inusual.




También parecen haber outliers en otras variables.




En particular, estos son los datos atípicos de HI

```
> boxplot.stats(data$HI)$out #Mostrar los datos atípicos
 [1]  9185.28  9350.13  9592.99  9249.13  9649.01  9174.83
 [7]  9935.07  9129.66 10000.00  9819.56  9130.73  9588.09
[13]  9978.49  9986.32
```

Estos los de las tarifas (FARE)

```
> boxplot.stats(data$FARE)$out #Mostrar los datos atípicos
[1] 367.72 374.40 374.40 374.40 402.02
```

Y estos son los de la variable PAX

```
> boxplot.stats(data$PAX)$out #Mostrar los datos atípicos
 [1] 30877 30877 34113 34113 32824 54429 51358 51358 29771 29771
[11] 29137 29137 48642 48642 48642 66820 66820 66820 73892 73892
[21] 73892 73892 73892 73892 40159 40159 40159 60435 60435 60435
[31] 51122 51122 51122 43884 54990 54990 54990 35471 27906 43671
[41] 41492 41492 41492 28988 37715 37715 27713 27713 27713 27713
[51] 63690 63690 63690 63690 63690 63690 34324 34324 34324
>
```

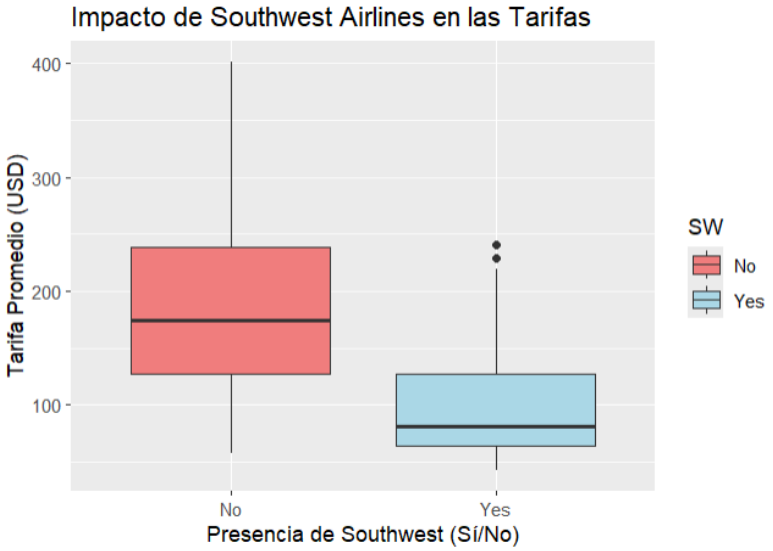
Si revisamos de nuevo la salida de la función summary() (esta vez con las variables como factores) podemos notar que realmente muchos "outliers" de PAX destacan porque superan el promedio, pero considero que no debe ser algo preocupante ya que sólo son viajes que superaron al promedio, y de hecho, no son tan poco comunes porque son alrededor de 60 viajes que fueron más caros que el promedio (representan un 9.4% del total de observaciones).

```
> summary(data)
    S_CODE             S_CITY             E_CODE             E_CITY             COUPON           NEW           VACATION    SW
 Length:638         Length:638         Length:638         Length:638         Min.   :1.000   Min.   :0.000   No :468   No :444
 Class :character   Class :character   Class :character   Class :character   1st Qu.:1.040   1st Qu.:3.000   Yes:170   Yes:194
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :1.150   Median :3.000
                                                                            Mean   :1.202   Mean   :2.754
                                                                            3rd Qu.:1.298   3rd Qu.:3.000
                                                                            Max.   :1.940   Max.   :3.000
       HI            S_INCOME        E_INCOME         S_POP            E_POP                SLOT              GATE
 Min.   : 1230   Min.   :14600   Min.   :14600   Min.   :  29838   Min.   : 111745   Controlled:182   Constrained:124
 1st Qu.: 3090   1st Qu.:24706   1st Qu.:23903   1st Qu.:1862106   1st Qu.:1228816   Free      :456   Free       :514
 Median : 4208   Median :28637   Median :26409   Median :3532657   Median :2195215
 Mean   : 4442   Mean   :27760   Mean   :27664   Mean   :4557004   Mean   :3194503
 3rd Qu.: 5481   3rd Qu.:29694   3rd Qu.:31981   3rd Qu.:7830332   3rd Qu.:4549784
 Max.   :10000   Max.   :38813   Max.   :38813   Max.   :9056076   Max.   :9056076
    DISTANCE          PAX             FARE
 Min.   : 114.0   Min.   : 1504   Min.   : 42.47
 1st Qu.: 455.0   1st Qu.: 5328   1st Qu.:106.29
 Median : 850.0   Median : 7792   Median :144.60
 Mean   : 975.7   Mean   :12782   Mean   :160.88
 3rd Qu.:1306.2   3rd Qu.:14090   3rd Qu.:209.35
 Max.   :2764.0   Max.   :73892   Max.   :402.02
```

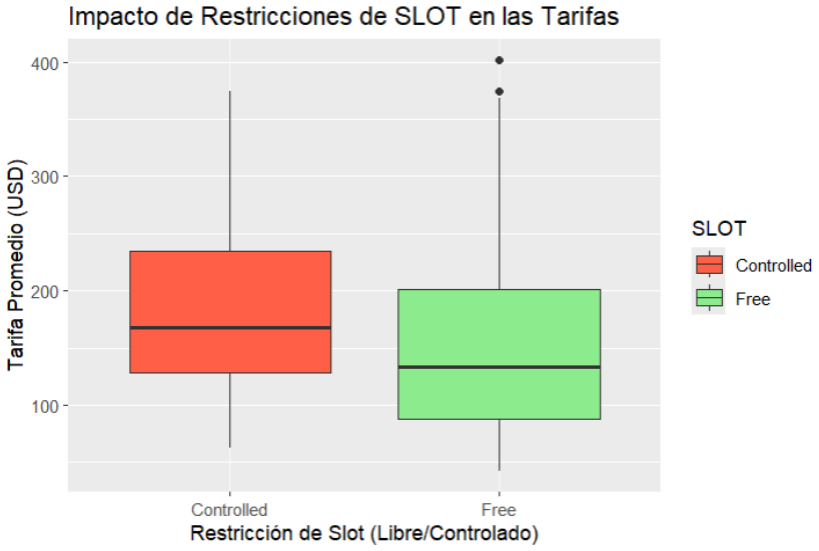
Con respecto a los outliers de FARE creo que podríamos eliminarlos si fuera necesario. Con los outliers de HI considero que no son tan atípicos como para eliminarlos, al menos de momento.

Para finalizar con el análisis exploratorio de datos hice unos ploteos de boxplots para observar el impacto de SW en las tarifas. Los resultados convencen sobre que la aerolínea SW tiene un impacto considerable en la reducción de los precios, lo cual es coherente con el texto del principio del PDF, donde se menciona que la alta congestión de aerolíneas se daba por una falta de competencia, hasta que se liberó la regulación y surgió más competencia con aerolíneas como SW.

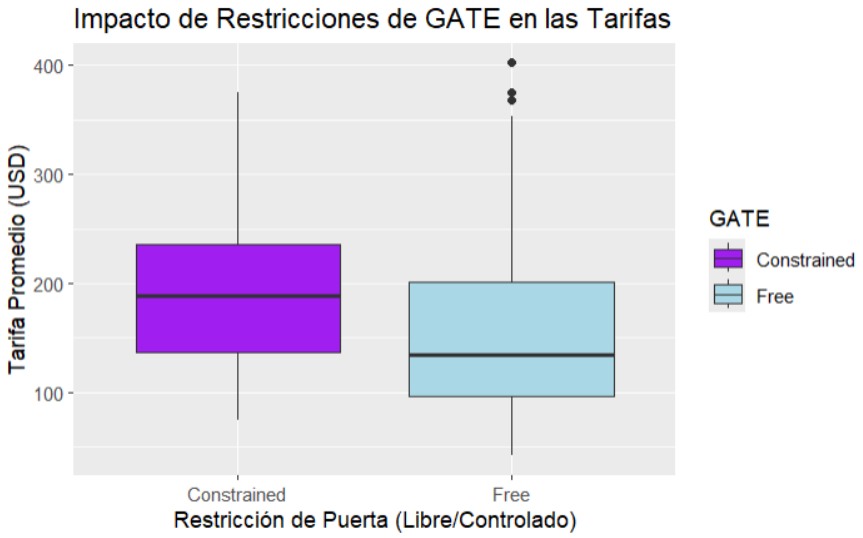
En este plot se puede ver que en definitiva, los vuelos de SW son más económicos que el resto. También se observa que hay menos variabilidad en las tarifas en las rutas donde Southwest opera, aunque hay algunos valores atípicos.




Los aeropuertos que tienen restricciones (Controlled) tienen tarifas un poco más caras, en comparación con los que no tienen restricciones (Free). Hay más dispersión en las tarifas de aeropuertos sin restricciones de slot, lo que sugiere una mayor variabilidad en los precios cuando no hay control en el uso de los slots. Y aunque la diferencia entre cajas no está tan marcada como en el plot anterior, se sugiere que la congestión puede impactar en el precio de las tarifas.



Aquí otra vez vemos que la diferencia no está tan marcada, pero sí lo suficiente como para decir que en los aeropuertos con restricciones de puerta ("Constrained") las tarifas son más altas en comparación con los aeropuertos donde las puertas están libres ("Free"). También, hay una mayor presencia de datos atípicos en los aeropuertos sin restricciones lo que indica que las tarifas pueden ser más variables en estos aeropuertos. Podemos interpretarlo como que las restricciones en las puertas limitan la competencia y reducen la disponibilidad de vuelos, lo que lleva a un aumento en los precios.

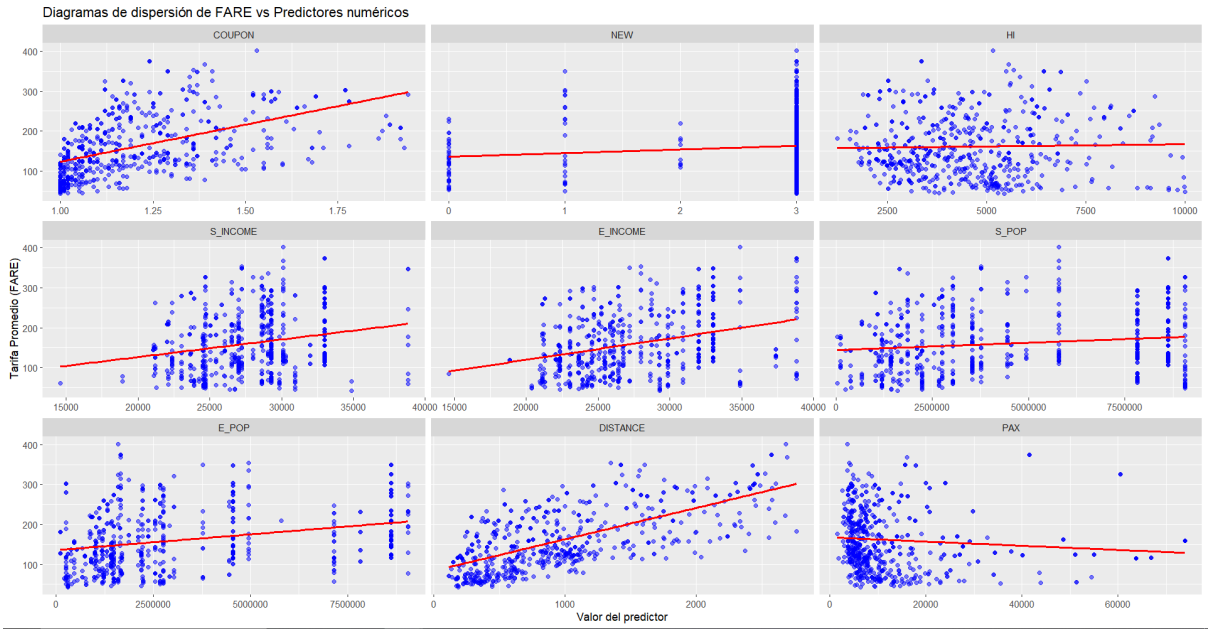



Ahora sí, empecemos con los incisos.

**A)** La tabla de correlación ya se creó en el EDA general (página 6), incluso se hizo un mapa de calor, por lo que iré directo a los diagramas de dispersión. Para empezar grafique primero los diagramas de aquellas variables con una mayor correlación con FARE (DISTANCE, PAX, COUPON y HI), pero para no llenar de capturas el reporte mejor hice un ploteo facetado que incluye todas las variables numéricas.

Ahora lo bueno, la interpretación de la captura. Podemos notar que a medida que el número de cupones aumenta (escalas en el vuelo), la tarifa promedio también sube, lo que es una correlación positiva. Por ejemplo, un vuelo sin escalas (1 cupón) es significativamente más barato que uno con más paradas, esto podría explicar por qué Southwest Airlines y otras aerolíneas de bajo costo apostaron por vuelos directos. También, la correlación entre distancia y tarifa es altísima, esto sugiere que el costo operativo (combustible, tripulación, mantenimiento) son el principal factor del precio, así que no importa si el vuelo es entre ciudades ricas o pobres, si la distancia es larga, el boleto será caro (con y sin cupón). A diferencia de la distancia, el número de

pasajeros en una ruta parece estar inversamente relacionado con la tarifa. Si una ruta es popular y hay suficiente competencia, las tarifas tienden a ser más bajas. Y sobre HI, parece que un mercado muy concentrado no necesariamente implica tarifas más altas, quizá porque los pasajeros tienen otras opciones en aquellas distribuciones más concentradas ( o sea, más competencia).




Ahora con esto en mente, ¿Cuál parece ser el mejor predictor individual de FARE?, pues DISTANCE. En primera porque su diagrama de dispersión es el más concentrado a lo largo de la línea roja y segundo porque es la variable con una mayor correlación con FARE.

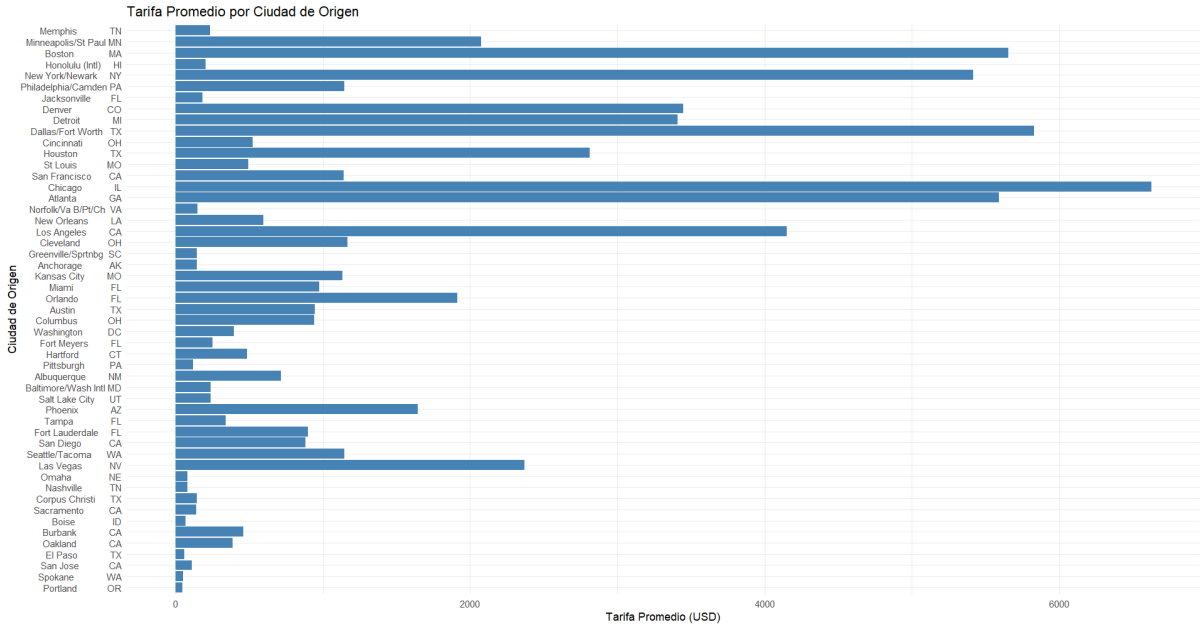
A manera de repaso:

- DISTANCE (cor. de 0.6700)
- COUPON (cor. de 0.4965)
- E_INCOME (cor. de 0.3260)

| Rango de correlación | Interpretación |
|---|---|
| $\pm$ 1.0 a $\pm$ 0.7 | Correlación fuerte |
| $\pm$ 0.7 a $\pm$ 0.5 | Correlación moderada |
| $\pm$ 0.5 a $\pm$ 0.3 | Correlación débil |
| $\pm$ 0.3 a 0.0 | Correlación muy débil o insignificante |

**B)** Para resolver este inciso me surgió la duda de si incluir S_CITY y E_CITY sería buena idea, pues, pese a que tienen muchas variables podría decirnos algo sobre las ciudades más visitadas, pese a que en el PDF se dice que no las incluyamos. Si ciertas ciudades tienen tarifas promedio significativamente diferentes, entonces podría valer la pena incluirlos. Pero si S_CITY y E_CITY tienen demasiadas categorías distintas (tantas como el número de ciudades en el dataset) es posible que no sean útiles porque dispersarían mucho los datos. En cambio, podrían ser útiles si agrupamos ciudades por características compartidas, como regiones geográficas (Norte, Sur, Este y Oeste); por el tamaño de la ciudad o por tráfico aeroportuario, pero este último sería redundante con la variable HI.

Para ver si vale la pena incluir esas variables calcule la tarifa promedio por ciudad de origen y destino (S_CITY y E_CITY), y luego lo grafique.




Viendo este plot me doy cuenta de que agrupar las ciudades por región puede no ser lo mejor porque algunas ciudades dentro de la misma región tienen tarifas muy distintas. Un enfoque más acertado es utilizar los datos de nuestro dataset para crear criterios de agrupación. La idea es agrupar las ciudades en:

- Hubs nacionales e internacionales.
- Aeropuertos nacionales y regionales.
- Aeropuertos low-cost.
- Aeropuertos secundarios/ de conectividad.
- Aeropuertos de tráfico local.
- Aeropuerto mediano regional.

Para ello me basaré en las siguientes variables:

- **DISTANCE**: Los vuelos de larga distancia suelen ser internacionales o transcontinentales. y los vuelos de corta distancia son generalmente domésticos y regionales.
- **S_POP** y **E_POP**: Las ciudades con población alta pueden ser grandes hubs con mucho tráfico. Y las ciudades más pequeñas pueden estar más orientadas a mercados regionales o low-cost.
- **HI**: Si el HI es alto, hay poca competencia, lo que puede indicar un hub dominante. Si es bajo, hay mucha competencia, típico de aeropuertos con aerolíneas low-cost.
- **SW**: Si está presente, la ciudad podría estar asociada con mercados de bajo costo. Si no está presente, la ciudad podría ser más dominada por aerolíneas tradicionales.

Por lo tanto, los criterios se evaluarán de la siguiente forma:

| Categoría | Criterios |
|---|---|
| Hubs nacionales e internacionales | Ciudades con alta población (S_POP > 5,000,000 o E_POP > 5,000,000).<br><br>Vuelos de larga distancia (DISTANCE > 800 millas) para evitar incluir vuelos cortos.<br><br>Un índice HI>5000 sugiere dominio de pocas aerolíneas. |
| Aeropuertos nacionales y regionales | Ciudades con población media (S_POP entre 1,000,000 y 5,000,000).<br><br>Vuelos de media distancia (DISTANCE > 500 millas). |
| Aeropuertos low-cost | Ciudades con presencia de aerolíneas de bajo costo o alta competencia en el mercado (SW == "Yes" o HI < 4000). |
| Aeropuertos secundarios o de conectividad | HI entre 3000 y 6000, indicando competencia moderada.<br><br>Vuelos de distancia media (DISTANCE < 1500) |

| | |
|---|---|
| Aeropuertos de tráfico local | Vuelos de corta distancia (DISTANCE < 600 millas) para incluir más aeropuertos.<br><br>HI > 6000, indica un mercado dominado por pocas aerolíneas. |
| Aeropuertos medianos regionales | Ciudades intermedias (S_POP entre 500,000 y 2,000,000).<br><br>Vuelos de distancia regional (DISTANCE entre 250 y 1000 millas). |
| Otros | No cumple con ningún criterio anterior |

Bueno, llegado a este punto hice varios ajustes en los criterios (originalmente planteé otros criterios que fui afinando hasta obtener los que están arriba en la tabla). Caí en cuenta de que podía usar árboles para clasificar mejor y así lo hice, el tema es que gracias a los árboles noté que podía modificar algunas cosas pero pareciera que estoy en un ciclo interminable de refinamiento porque llevo mucho tiempo en esto así que lo dejaré con los criterios actuales para poder avanzar en los demás incisos. De todas formas el árbol me demostró que aunque hay cosas que mejorar no está nada mal mi planteamiento actual.

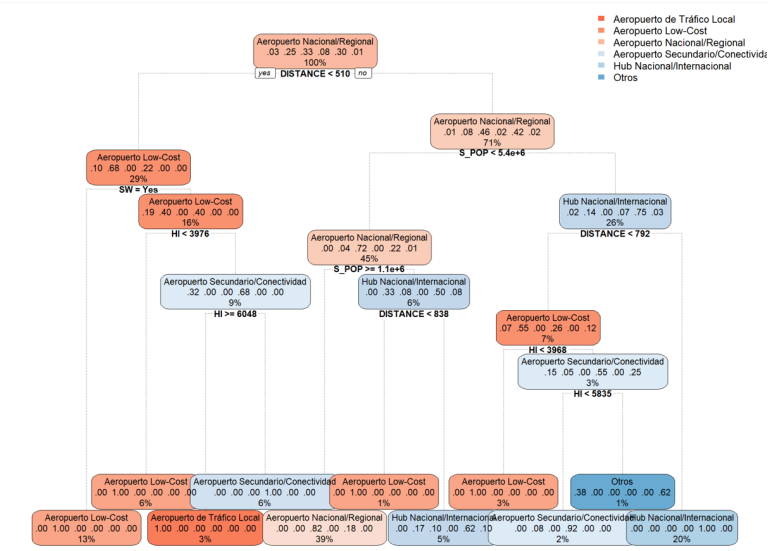
Así quedó la distribución de los aeropuertos:

```
> #Distribución de la clasificación
> table(data$CITY_TYPE)

       Aeropuerto de Tráfico Local                Aeropuerto Low-Cost       Aeropuerto Nacional/Regional 
                                22                                159                                208 
Aeropuerto Secundario/Conectividad         Hub Nacional/Internacional                              Otros 
                                51                                190                                  8 
```

Y estos fueron los resultados de mi árbol:

```
> summary(tree_model)
Call:
rpart(formula = CITY_TYPE ~ DISTANCE + S_POP + HI + SW, data = data,
    method = "class")
  n= 638

          CP nsplit rel error    xerror       xstd
1 0.29069767      0 1.0000000 1.0000000 0.02753513
2 0.05348837      2 0.4186047 0.4279070 0.02661081
3 0.04651163      3 0.3651163 0.3813953 0.02567038
4 0.04418605      5 0.2720930 0.3372093 0.02461663
5 0.03488372      6 0.2279070 0.2581395 0.02226835
6 0.02325581      7 0.1930233 0.2023256 0.02015843
7 0.01627907      8 0.1697674 0.1767442 0.01902813
8 0.01162791      9 0.1534884 0.1767442 0.01902813
9 0.01000000     10 0.1418605 0.1837209 0.01934820

Variable importance
DISTANCE    S_POP       HI       SW 
      41       30       24        6 
```

```
> #Evaluación - manual vs predicción
> table(data$CITY_TYPE, data$CITY_TYPE_PRED)

                                Aeropuerto de Tráfico Local Aeropuerto Low-Cost Aeropuerto Nacional/Regional
Aeropuerto de Tráfico Local                              19                   0                           0
Aeropuerto Low-Cost                                       0                 153                           0
Aeropuerto Nacional/Regional                              0                   0                         205
Aeropuerto Secundario/Conectividad                        0                   0                           0
Hub Nacional/Internacional                                0                   0                          46
Otros                                                     0                   0                           0

                                Aeropuerto Secundario/Conectividad Hub Nacional/Internacional Otros
Aeropuerto de Tráfico Local                                      0                         0     3
Aeropuerto Low-Cost                                              1                         5     0
Aeropuerto Nacional/Regional                                     0                         3     0
Aeropuerto Secundario/Conectividad                              51                         0     0
Hub Nacional/Internacional                                       0                       144     0
Otros                                                            0                         3     5
```




Bueno, luego de tanto rollo vayamos directo a lo que se solicita en el inciso.

```
Distribución de CITY_TYPE :

       Aeropuerto de Tráfico Local
                              3.45
              Aeropuerto Low-Cost
                             24.92
      Aeropuerto Nacional/Regional
                             32.60
Aeropuerto Secundario/Conectividad
                              7.99
        Hub Nacional/Internacional
                             29.78
                             Otros
                              1.25

Distribución de VACATION :

   No   Yes
73.35 26.65

Distribución de SW :

   No   Yes
69.59 30.41

Distribución de SLOT :

Controlled       Free
     28.53      71.47

Distribución de GATE :

Constrained        Free
      19.44       80.56
```

```
Tarifa Promedio según CITY_TYPE :
                           CITY_TYPE Mean_FARE
1        Aeropuerto de Tráfico Local  165.3786
2               Aeropuerto Low-Cost  100.1357
3      Aeropuerto Nacional/Regional  163.2948
4 Aeropuerto Secundario/Conectividad  156.2398
5         Hub Nacional/Internacional  209.8051
6                             Otros  160.3613

Tarifa Promedio según VACATION :
  VACATION Mean_FARE
1       No  173.5525
2      Yes  125.9809

Tarifa Promedio según SW :
   SW Mean_FARE
1  No 188.18279
2 Yes  98.38227

Tarifa Promedio según SLOT :
        SLOT Mean_FARE
1 Controlled  186.0594
2       Free  150.8257

Tarifa Promedio según GATE :
         GATE Mean_FARE
1 Constrained   193.129
2        Free   153.096
```

Al parecer el mejor predictor SW (si omitimos los primeros 4 como se pidió en las instrucciones), aunque no es el único predictor fuerte. Las categorías Hub nacional/internacional y Low-Cost de CITY_TYPE también tienen mucho impacto en las tarifas, uno sobre las tarifas caras y otro sobre las más baratas, por lo que uno puede predecir o hacerse a la idea del precio de una tarifa a partir de CITY_TYPE. Esto lo podemos saber porque si calculamos la diferencia entre el máximo y el mínimo de cada variable para medir su impacto en las tarifas, la diferencia de Hub y low-cost es la más grande.

```
> print(fare_diff)
$CITY_TYPE
[1] 109.6694

$VACATION
[1] 47.57162

$SW
[1] 89.80052

$SLOT
[1] 35.23372

$GATE
[1] 40.03308
```

C) Bueno, conservé la variable CITY_TYPE y la convertí a dummy. Luego hice el modelo por pasos y descartó las variables CITY_TYPE.Otros y CUPON.

```
> summary(model_step)

Call:
lm(formula = FARE ~ `CITY_TYPE.Aeropuerto Low-Cost` + `CITY_TYPE.Aeropuerto Nacional/Regional` +
    `CITY_TYPE.Aeropuerto Secundario/Conectividad` + `CITY_TYPE.Hub Nacional/Internacional` +
    VACATION.Yes + SW.Yes + SLOT.Free + GATE.Free + NEW + HI +
    S_INCOME + E_INCOME + S_POP + E_POP + DISTANCE + PAX, data = train)

Residuals:
   Min     1Q Median     3Q    Max
-87.80 -22.13  -1.70  20.55 126.72

Coefficients:
                                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                                 -7.614e+00  2.559e+01  -0.298 0.766182
`CITY_TYPE.Aeropuerto Low-Cost`              2.161e+01  9.434e+00   2.290 0.022433 *
`CITY_TYPE.Aeropuerto Nacional/Regional`     2.879e+01  9.107e+00   3.161 0.001668 **
`CITY_TYPE.Aeropuerto Secundario/Conectividad` 3.370e+01  9.970e+00   3.380 0.000783 ***
`CITY_TYPE.Hub Nacional/Internacional`       3.201e+01  9.969e+00   3.211 0.001409 **
VACATION.Yes                                -3.812e+01  4.164e+00  -9.154  < 2e-16 ***
SW.Yes                                      -4.075e+01  4.431e+00  -9.198  < 2e-16 ***
SLOT.Free                                   -1.590e+01  4.273e+00  -3.721 0.000221 ***
GATE.Free                                   -1.683e+01  4.453e+00  -3.780 0.000176 ***
NEW                                         -5.007e+00  2.104e+00  -2.379 0.017719 *
HI                                           9.550e-03  1.179e-03   8.099 4.33e-15 ***
S_INCOME                                     1.221e-03  5.750e-04   2.123 0.034216 *
E_INCOME                                     1.358e-03  4.248e-04   3.198 0.001472 **
S_POP                                        3.055e-06  8.422e-07   3.627 0.000316 ***
E_POP                                        4.175e-06  9.402e-07   4.440 1.11e-05 ***
DISTANCE                                     7.337e-02  3.583e-03  20.479  < 2e-16 ***
PAX                                         -7.625e-04  1.537e-04  -4.960 9.69e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 35.41 on 496 degrees of freedom
Multiple R-squared:  0.792,     Adjusted R-squared:  0.7853
F-statistic:   118 on 16 and 496 DF,  p-value: < 2.2e-16
```

```
> cat("Variables eliminadas:", setdiff(names(coef(reg_lin)), names(coef(model_step))), "\n")
Variables eliminadas: CITY_TYPE.Otros COUPON
```

Aquí vemos factores clave como que las rutas más largas aumentan tarifas, más pasajeros las reducen, aeropuertos congestionados elevan precios y la presencia de Southwest (-$40.75) o rutas turísticas disminuye costos. Las ciudades grandes y con altos ingresos tienen tarifas más altas. Evaluar error en prueba.

Luego, la búsqueda exhaustiva resultó en un modelo más simple (11 predictores en lugar de 16), descartando tipos de aeropuerto y los ingresos de la ciudad de origen (S_INCOME). Sin embargo, la regresión por pasos retuvo más información contextual sobre la conectividad de los aeropuertos y la competencia en el mercado. Las variables en común fueron: VACATION.Yes, SW.Yes, SLOT.Free y GATE.Free, HI, E_INCOME, S_POP y E_POP, DISTANCE y PAX.

Si comparamos la precisión de los modelos obtenemos lo siguiente:

```
> print(paste("Regresión por pasos - RMSE:", round(rmse_step, 2), "| MAE:", round(mae_step, 2)))
[1] "Regresión por pasos - RMSE: 34.37 | MAE: 28.12"
> print(paste("Búsqueda exhaustiva - RMSE:", round(rmse_ex, 2), "| MAE:", round(mae_ex, 2)))
[1] "Búsqueda exhaustiva - RMSE: 33.78 | MAE: 27.39"
```

Al parecer la búsqueda exhaustiva es un poco mejor que la regresión por pasos.

Si hacemos una predicción con el modelo exhaustivo para los datos dados en el subinciso ***v*** obtenemos lo siguiente.

Nota: como el modelo exhaustivo incluyó CITY_TYPE.Otro como variable asumí que no era de ese tipo por la distancia.

```
> tarifa_pred
       1
253.2084
> print(paste("Tarifa promedio predicha:", round(tarifa_pred, 2), "USD"))
[1] "Tarifa promedio predicha: 253.21 USD"
```

Si esta vez cambiamos el valor de SW de 0 a 1 (lo que significa que SW cubre la ruta) obtenemos lo siguiente:

```
> tarifa_predicha_SW
       1
209.163
> print(paste("Tarifa promedio con Southwest:", round(tarifa_predicha_SW, 2), "USD"))
[1] "Tarifa promedio con Southwest: 209.16 USD"
```

Lo que significa que la reducción del precio es igual a la diferencia de ambas tarifas.

```
> #Calcular reducción en la tarifa
> reduccion_tarifa <- tarifa_pred - tarifa_predicha_SW
> reduccion_tarifa
       1
44.04538
```

Respondiendo al inciso ***vii***. Las variables (o factores) que no podemos considerar en una predicción para un aeropuerto nuevo son: **PAX**, **HI**, **FARE**. El número de pasajeros en la ruta no lo podemos conocer antes de la apertura porque tienen que realizarse los vuelos para poder saber cuántos pasajeros se llevan (en todo caso se podría hacer una estimación, pero sólo eso). Sobre HI sucede algo curioso, pues, en teoría nosotros de antemano podemos saber cómo está el HI de San Francisco a Chicago, pero esto es una observación incompleta debido a que si el aeropuerto aún no está operativo no podemos saber cómo dicho aeropuerto impacta en el mercado y por lo tanto en esa ruta. Y finalmente, sobre FARE no podemos generar un promedio de las tarifas si no hay un histórico de las mismas.

Por otro lado, las variables que sí podemos considerar son: DISTANCE, **S_POP**, **E_POP** (Población en ciudades de origen/destino), **S_INCOME**, **E_INCOME** (Ingreso promedio en ciudades), **SW.Yes**, **SLOT.Free**, **GATE.Free** (Restricciones del aeropuerto) y **VACATION.Yes** (ya que de antemano uno puede especificar si la ruta es principalmente turística o no). Cada una de esas variables forman parte de la logística de la construcción del aeropuerto, excepto SW que como tal no forma parte de la logística de construcción pero sí de la logística de planeación (como una estrategia de expansión, por ejemplo). Al igual, la distancia, SLOT y GATE son una combinación de ambas logísticas (de construcción y de planeación).

***viii)*** El modelo mantuvo las variables que desde un inicio destacaban por su relevancia (al menos las que se pueden considerar para un aeropuerto nuevo).

```
> best_model_vars_reducido
  (Intercept)      DISTANCE      E_INCOME     SLOT.Free     GATE.Free        SW.Yes  VACATION.Yes
111.282606576   0.073330294   0.001412916 -16.790696268 -25.127556769 -49.787891556 -50.946116425
```

**ix)** Al parecer el modelo reducido predice una tarifa promedio de $253.35 USD, lo cual es muy cercano a la predicción del modelo completo ($253.21 USD), incluso pese a que no se considera a HI en el modelo (ya que recordemos que el modelo que hice no consideraba a HI como un factor que podemos saber antes de abrir el aeropuerto). Seguramente la precisión similar a la del modelo completo se deba a que se conservan muchas de las variables más importantes.

```
> tarifa_pred_reducida
[1] 253.3519
> print(paste("Tarifa promedio predicha con modelo reducido:", round(tarifa_pred_reducida, 2), "USD"))
[1] "Tarifa promedio predicha con modelo reducido: 253.35 USD"
```

**x)** Como se mencionó anteriormente, el modelo reducido se acerca al modelo completo. Aunque el modelo reducido no tiene algunas variables importantes sigue teniendo un error relativamente bajo. De cara a la reciente apertura del aeropuerto, seguramente la precisión mejore rápidamente con la acumulación de un histórico de

las tarifas. Y comparando ambas precisiones, el modelo completo, en promedio, tiene predicciones con $4.13 menos de error que el modelo reducido.

```
> print(paste("Modelo original (iii) - RMSE:", round(rmse_ex, 2), "| MAE:", round(mae_ex, 2)))
[1] "Modelo original (iii) - RMSE: 33.78 | MAE: 27.39"
> print(paste("Modelo reducido (viii) - RMSE:", round(rmse_ex_reducido, 2), "| MAE:", round(mae_ex_reducido, 2)))
[1] "Modelo reducido (viii) - RMSE: 38.51 | MAE: 31.52"
```

**D)** Si cambiamos nuestro enfoque de análisis pasando de la predicción de tarifas en nuevas rutas al análisis del impacto de SW en el mercado pasaría de usar sólo modelos de regresión a usar series de tiempo para medir cómo cambiaron las cosas en el mercado antes y después de la aparición de SW. Realmente con todo el trabajo "extra" que hice me fui dando cuenta del impacto que tiene SW en el mercado, por ejemplo, en la categorización de CITY_TYPE noté que las aerolíneas low-cost tienen una presencia considerable en el mercado, de hecho, son la tercera categoría con más vuelos, entonces eso nos ayuda a darnos una idea del impacto que tiene en el mercado. Además, en los modelos también se consideraron dichas variables que consideran el mercado en general (no sólo SW), en este sentido, la regresión por pasos me pareció más completa ya que consideraba más variables de mercado que la búsqueda exhaustiva, pero como el objetivo era predecir tarifas resultó ser más adecuada la búsqueda. Además, en la parte en que se cambió únicamente el estado de SW.Yes también ilustró el impacto de SW en el mercado.

Como tal, hay mucha tela de la que cortar en este ejercicio, se puede abordar de múltiples formas y se nota en que tuve que usar métodos adicionales (árboles) para figurarme una mejor solución de las cosas, de hecho, bien pude hacer clustering al inicio para identificar grupos clave. Si bien las series de tiempo serían de mucha ayuda para analizar el mercado, no son la única forma de hacerlo. Lo que cambiaría sería la selección de modelos, de variables de importancia y obviamente la inclusión de nuevas técnicas de análisis como series de tiempo.

## Situación financiera de los bancos

A) El modelo de regresión solicitado da como resultado lo siguiente:

```
> summary(model_bank)

Call:
glm(formula = Financial.Condition ~ TotLns.Lses.Assets + TotExp.Assets,
    family = binomial, data = data_bank)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -14.721      6.675  -2.205   0.0274 *
TotLns.Lses.Assets   8.371      5.779   1.449   0.1474
TotExp.Assets       89.834     47.781   1.880   0.0601 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 27.726  on 19  degrees of freedom
Residual deviance: 13.148  on 17  degrees of freedom
AIC: 19.148

Number of Fisher Scoring iterations: 6
```

Por lo que las ecuaciones estimadas se verían así:

**i)**

$$log(\frac{P(Débil)}{1-P(Débil)}) = \beta_0 + \beta_1 * \ TotLns.Lses.Assets + \beta_2 * TotExp.Assets$$

Sustituyendo valores quedaría como:

$$log(\frac{P(Débil)}{1-P(Débil)}) = -\ 14.721 + 8.371 * \ TotLns.Lses.Assets + 89.834 * TotExp.Assets$$

**ii)** $\frac{P(Débil)}{1-P(Débil)} = e^{\beta_0+\beta_1*TotLns.Lses.Assets+\beta_2*TotExp.Assets}$

Sustituyendo valores quedaría como:

$$\frac{P(Débil)}{1-P(Débil)} = e^{-14.721+8.371*\ TotLns.Lses.Assets+\ 89.834*TotExp.Assets}$$

**iii)** $P(Débil) = \frac{1}{1+e^{-(\beta_0+\beta_1*TotLns.Lses.Assets+\beta_2*TotExp.Assets)}}$

Sustituyendo valores quedaría como:

$$P(Débil) = \frac{1}{1+e^{-(-14.721+8.371*\ TotLns.Lses.Assets+\ 89.834*TotExp.Assets)}}$$

**B)** Para la mala fortuna de nuestros banqueros el modelo clasificó al nuevo banco como financieramente débil. La probabilidad obtenida de que será financieramente débil es del 54% (por poco superó el punto de corte de 0.5). Aún así, la probabilidad no es tan grande como para tirar por la borda todo, de hecho, nuestro logit de 0.1835 indica que el banco tiene una **ligera** tendencia a ser clasificado como débil. Y nuestros odds dicen que la probabilidad de que el banco sea débil es aproximadamente 1.2 veces la probabilidad de que sea fuerte.

```
> cat("Logit:", round(logit_val, 4), "\n")
Logit: 0.1835
> cat("Odds:", round(odds, 4), "\n")
Odds: 1.2014
> cat("Probabilidad de ser débil:", round(weak_prob, 4), "\n")
Probabilidad de ser débil: 0.5458
> clasificacion
(Intercept)
    "Débil"
```

**C)** Para resolver lo que se pide podemos partir de una de las fórmulas de arriba:

$$Logit(\pi) = log(\frac{P(Débil)}{1-P(Débil)}) = log(\frac{0.6}{1-0.6}) = 0.4055$$

Siendo honesto el resultado no lo obtuve manualmente, lo calculé con R de la siguiente forma:

```
> #El nuevo umbral de probabilidad (p) es 0.6, así que:
> new_prob<-0.6
> nuevo_umbral_logit <- log(new_prob / (1 - new_prob))
> cat("Nuevo umbral de probabilidad:", round(new_prob, 4), "\n")
Nuevo umbral de probabilidad: 0.6
> cat("Umbral equivalente en logit:", round(nuevo_umbral_logit, 4), "\n")
Umbral equivalente en logit: 0.4055
```

Bueno, como dije sería un punto de partida porque el decir que 0.5 o 0.6 serán puntos de corte es dar valores arbitrarios, así que para encontrar el mejor umbral mejor usemos la curva ROC para ver cómo varían la sensibilidad y especificidad del modelo en diferentes umbrales de probabilidad sin tener que hacer manualmente cambios de corte. Observaremos el mejor umbral para el punto de corte y lo convertiremos a logit para tener una mejor interpretación con respecto a los datos de arriba.

```
> mejor_umbral_prob
  threshold
1 0.4546714
> mejor_umbral_logit
   threshold
1 -0.1818136
```

Según el índice de Youden el mejor corte para clasificar un banco como débil es 0.4547. Y en lenguaje logit (obtenido mediante la fórmula de log() de arriba), cuando el logit estimado para un banco es mayor o igual a -0.1818, entonces ese banco se clasificará como "débil".

**D)** Recordemos que nuestro coeficiente estimado para TotLns&Lses/Assets = $\beta_1$, o sea que equivale a lo siguiente:

```
> beta_1 <- coeficientes[2] #TotLns.Lses.Assets
> beta_1
TotLns.Lses.Assets
          8.37132
```

Así que para interpretarlo en términos de la probabilidad de ser financieramente débil otra vez hay que regresar a la fórmula de log().

$$log(\frac{P(Débil)}{1-P(Débil)}) = \beta_0 + \beta_1 * TotLns.Lses.Assets + \beta_2 * TotExp.Assets$$

Aquí $\beta_1$ indica el cambio de logit por cada aumento de TotLns.Lses.Assets, o sea, cuánto cambia logit cuando TotLns.Lses.Assets aumenta una unidad. Como tenemos $\beta_1 = 8.37132$ podemos decir que $\beta_1$ tiene un alto impacto positivo en la clasificación de bancos débiles (si $\beta_1$ es positivo significa que hay una mayor posibilidad de ser débil, y si es negativo al revés). Pero para poder aterrizarlo en términos de probabilidad hay que usar la función sigmoide de ***ii)***. Hice la mención de la fórmula log() para explicar por qué definí los aumentos de cambio en 0.1

```
> cambio <- 0.1
> logit_change <- beta_1 * cambio
> prob_change <- 1 / (1 + exp(-logit_change))
> cat("Cambio en la probabilidad por un incremento de 0.1 en TotLns.Lses.Assets:", round(prob_change, 4), "\n")
Cambio en la probabilidad por un incremento de 0.1 en TotLns.Lses.Assets: 0.6979
```

E) Para saber si debemos aumentar el valor de corte partiendo de 0.5 hice un arreglo que va de 0.3 a 0.8 en aumentos de 0.05 y luego calculé la sensibilidad y especificidad de cada corte para compararlos en una tabla.

```
> print(resultados)
   Umbral Sensibilidad Especificidad
1    0.30           1.0           0.7
2    0.35           1.0           0.7
3    0.40           1.0           0.8
4    0.45           1.0           0.9
5    0.50           0.9           0.9
6    0.55           0.9           0.9
7    0.60           0.8           0.9
8    0.65           0.8           0.9
9    0.70           0.8           0.9
10   0.75           0.6           0.9
11   0.80           0.6           0.9
```

Como podemos ver, si aumentamos el corte por encima de 0.5 la sensibilidad cae hasta 0.6, según la especificidad seguimos clasificando bien a los bancos fuertes, pero realmente nuestro objetivo es evitar clasificar erróneamente bancos débiles. Entonces, si bajamos el corte por debajo de 0.5 empieza a mejorar la detección de bancos débiles a costa de clasificar erróneamente algunos bancos fuertes. Dicho de forma burda, es como si al disminuir el corte dijeras con más facilidad que "x" banco es débil y por eso los detectas mejor porque justo ese comportamiento es similar a la detección de bancos fuertes cuando aumentamos el umbral por encima de 0.5. Eso no quiere decir que esté mal pero si es algo a considerar porque tampoco podemos disminuir el corte de manera deliberada porque si no el modelo va a categorizar cualquier banco como débil. Entonces, ¿se debe aumentar o reducir el corte de 0.5?, pues para penalizar el error más costoso (no detectar un banco débil) es mejor disminuir el corte, pero, ¿Qué tanto?, según la tabla entre 0.4 y 0.45 está bien, pero yo prefiero dar un corte específico porque este rango refuerza la validez del corte óptimo que calculamos con la curva ROC, **0.4547**.

## Conclusión

El segundo ejercicio no fue tan complicado como el primero debido a la naturaleza de los datos y a cómo se relacionan entre sí. Definitivamente el EDA del primer ejercicio me llevó mucho tiempo, y la inclusión (contraindicada) de las variables de ciudad también me tomó mucho tiempo. Aunque no aportaron explícitamente al resultado final si me dieron una perspectiva más amplia del problema ya que se me hizo más natural inferir cómo cambiaría el enfoque si quisiera analizar el impacto de SW en el mercado (de hecho, si hubiese seguido por ese camino el trabajo extra que hice para la variable CITY_TYPE hubiera sido clave).

El segundo problema se me hizo menos disfrutable, quizá porque no vi arder las billeteras de los corpos (corporativos). Independientemente de eso, fue agradable retomar la curva ROC para minimizar el error más costoso. Mientras que en el primer problema se exploraron enfoques para optimizar la predicción de tarifas, en el segundo se tomó en cuenta el impacto de los errores de clasificación en la toma de decisiones (hablando de decisiones, quizá podría implementarse un árbol de decisión para comparar los resultados con los de las instrucciones).

En fin, lo que me llevo es el ejercicio de exploración profunda de los datos, la selección adecuada de variables y la interpretación de los resultados para la aplicación de modelos predictivos en problemas del mundo real (y también unas cuantas desveladas con música de Cyberpunk 2077).