Friday, 27 March 2020

Student Name/ID Number	CELIS VASQUEZ SONIA PATRICIA
Unit Number and Title	**12: Data Analytics**
Academic Year	2020
Unit Tutor	Daniel González Martínez
Assignment Title	**Data Analytics: Prescriptive Analytics**
Issue Date	**February 2nd , 2020**
Submission Date	March 27th, 2020
IV Name & Date	Luis Ortiz

DATA ANALYTICS: PREDICTIVE ANALYTICS

A. Introduction to data analytics

1. Define, briefly, the following prescriptive analytic methods and indicate two examples of analytic techniques for each of them:

a. Optimization:

Se puede decir que la **Optimization** nos permite seleccionar la **MEJOR** solución por medio de algoritmos o soluciones matemáticas para la toma de decisiones con base a algún criterio especifico.

En este método debe existir un objetivo que nos permita diferenciar entre varias soluciones validas.

Se utilizan técnicas como: la programación lineal, la programación entera y la programación no lineal.

Examples:

Máximizar ó Minimizar.

Provisión de mercancia.

Evaluación de la forma de pago de un cliente para obtención de créditos.

Predicción de recursos para la fabricación de productos.

b. Decision analysis:

Se puede decir qué **Decisión Analysis** nos permite tomar distintas soluciones/acciones de forma sistemática en la toma de decisiones dependiendo de multiples y diversas reglas de negocio, lo que mejora su rendimiento.

Se utilizan técnicas basadas en reglas como son: los motores de inferencia, las tablas de puntuación, árboles de decisión.

Examples:

Previsión en él calculo del precio del seguro de vehículos.

Mantenimiento preventivos – cambio de piezas, rendimiento de equipos.

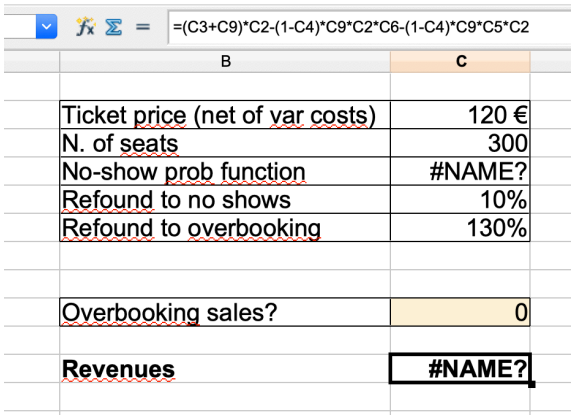
Recomendador de productos basados en las compras anteriores.

Analysis de sentimientos.

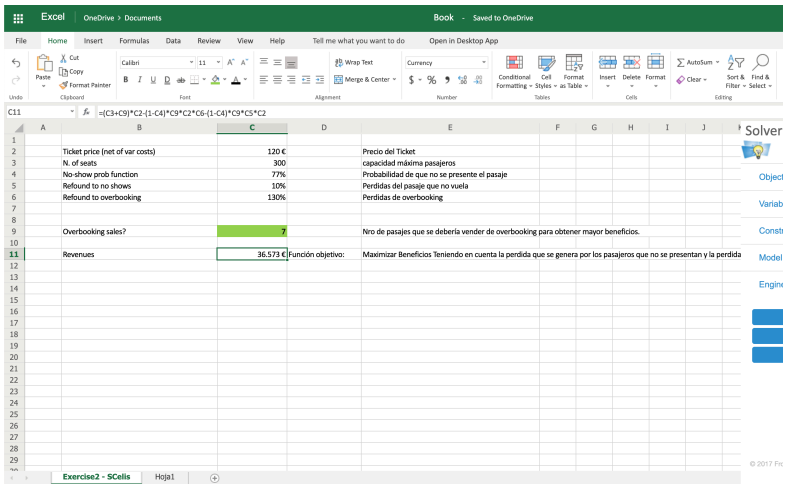
DATA ANALYTICS: PREDICTIVE ANALYTICS

2. Open the given Excel file (prescriptive_overbooking.xlsx).
   The airline company Avio noticed that in each flight there are several no-shows (passengers that had bought a ticket but didn't fly) and your boss wants to sell more tickets than the total number of seats to optimize revenues. You are asked to identify the number of extra tickets the company should sell to maximize revenues. All the variables and formulas are given in the Excel file. Use the Solver Add-in to maximize Total Revenues changing the variable Overbooking Tickets.

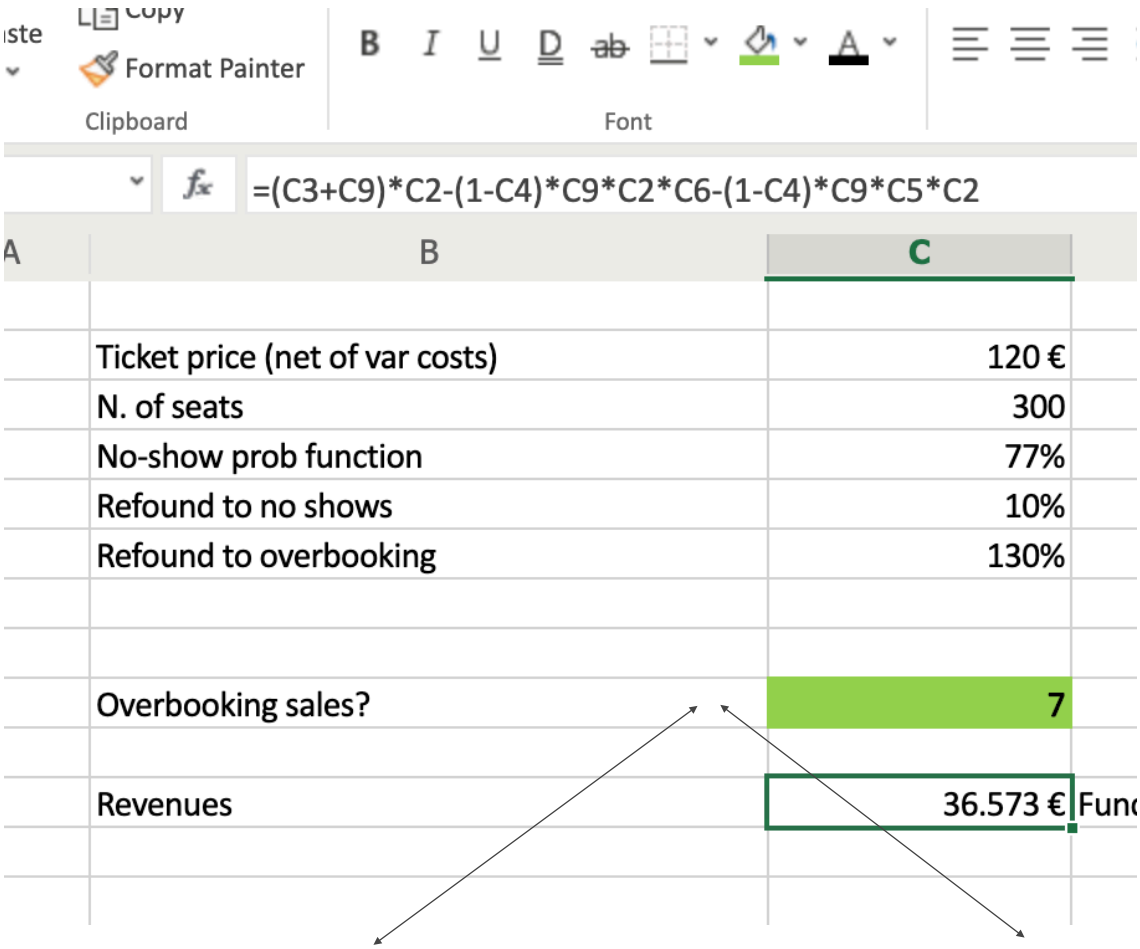
   a. - What is the number of Overbooking Tickets that maximizes revenues?



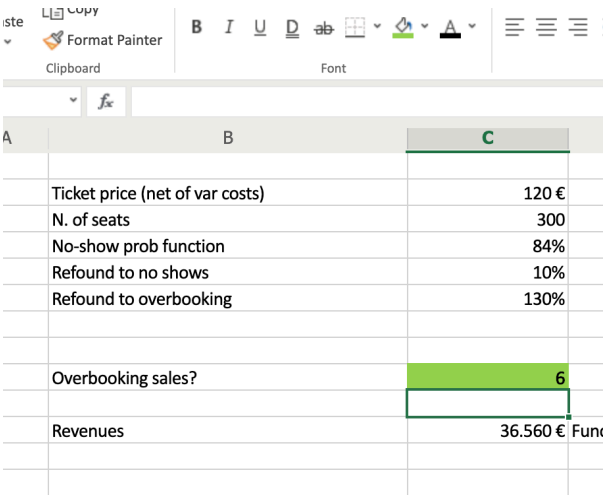
Teniendo en cuenta que nos están indicando el precio del ticket, la cantidad del pasaje, y los porcentajes de perdidas que están teniendo cuando el cliente compra el ticket y no se presenta al vuelo 10% y el porcentaje de perdidas generadas cuando hay exceso de reservas de tickets 130%, se plantea la necesidad de ajustar el número apropiado de tickets que se debería vender en overbooking para que podamos maximizar el total de los beneficios.

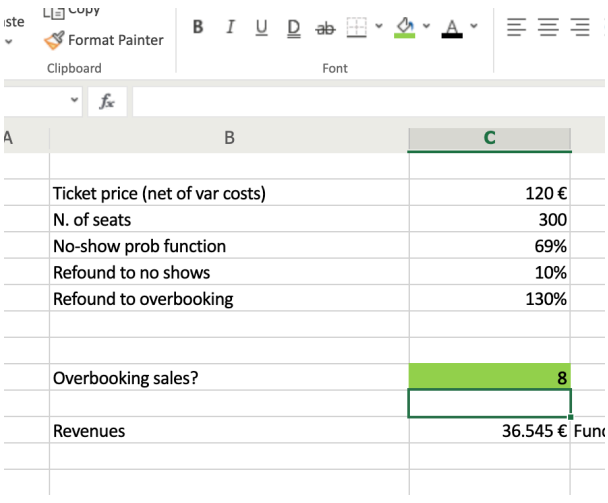


Podemos concluir que para obtener su mayor beneficio **36.573€** deberían vender **7** tickets de overbooking.



Si venden más ó menos tickets el beneficio y el porcentaje de probabilidad de que no se presenten al vuelo varian.





3. In the previous assignment you used a multiple linear regression to identify the impact of Avio's prices and competitors' prices on flight demand. You finally identified the regression function to estimate flight demand. Describe how you could use this function to create an optimization model:

    a. - What would be the objective of the optimization?

    b. - What is the decision variable?

    c. Can you imagine any constrain concerning the Pax (demand) variable?

– El objetivo de la optimización es optimizar el modelo que maximiza el beneficio en función de la demanda.

– La variable de decisión en este caso es el número de tickets que hay que permitir de overbooking, ya que es la variable que realmente determina los beneficios ya que genera un mayor porcentaje de perdida.

– Existen muchas limitaciones con respecto a la demanda, por ejemplo las huelgas de los controladores, incluso las crisis mundiales, guerras (en la actualidad) virales, que generan prohibiciones de desplazamientos y/o desplazamientos masivos, por lo tanto se debe optimizar para que ajuste la oferta a la demanda de forma adecuada y/o permita una mejor gestión del overbooking sin que afecte nuestros beneficios de manera considerable.

4. The CEO of Avio looked at satisfaction data in different months and claims that average satisfaction is significantly different in summer compared to winter. To verify his statement you take a sample of 30 clients who have travelled both in summer and in winter (in the following table). At 5% level of significance, test to see if the evidence supports the CEO's theory.

State the hypothesis in words, and perform a t-test to test whether the evidence supports the physician's theory, at the α = 5%. (Use a programming language or a data analytic tool.)

Perform the parametric t-test.

State the hypothesis: null hypothesis and alternative hypothesis.

Report normality test result using p-value.

The value of the test statistic is =

Write the conclusion using p-value.

Also comment on whether the evidence is statistically significant enough to support the physician's claim.

Perform the nonparametric signed rank test.

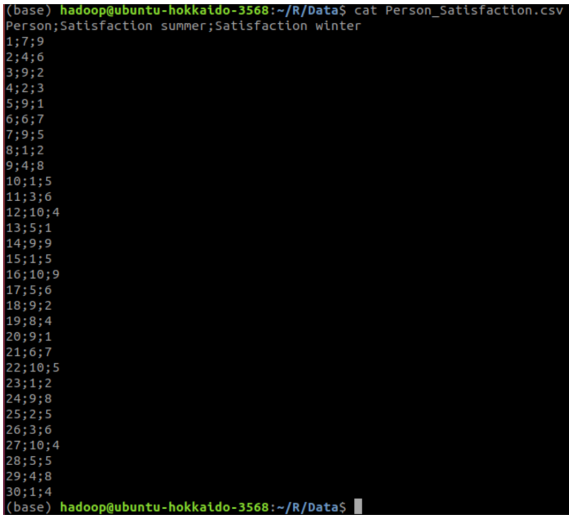
Value of the test statistic is =

Conclusion with p-value.

Also comment on whether the evidence is statistically significant enough to support the CEO's claim.

- He convertido la tabla en un fichero csv y he trabajado todo el ejercicio en R.

Visualización del fichero:



> satisfaction
> satisfaction

	Person	Satisfaction.summer	Satisfaction.winter
1	1	7	9
2	2	4	6
3	3	9	2
4	4	2	3
5	5	9	1
6	6	6	7
7	7	9	5
8	8	1	2
9	9	4	8
10	10	1	5
11	11	3	6
12	12	10	4
13	13	5	1
14	14	9	9
15	15	1	5
16	16	10	9
17	17	5	6
18	18	9	2
19	19	8	4
20	20	9	1
21	21	6	7
22	22	10	5
23	23	1	2
24	24	9	8
25	25	2	5
26	26	3	6
27	27	10	4
28	28	5	5
29	29	4	8
30	30	1	4

Lectura del fichero:

```
> satisfaction <- read.csv(file = "Person_Satisfaction.csv",
+ stringsAsFactors=FALSE,
+ strip.white=TRUE,
+ sep=";")
```

Definición de variables:

```
> names(satisfaction)
[1] "Person"              "Satisfaction.summer" "Satisfaction.winter"
> names(satisfaction)[2] = "summer"
> names(satisfaction)[3] = "winter"
> names(satisfaction)
[1] "Person" "summer" "winter"
> head(satisfaction)
  Person summer winter
1      1      7      9
2      2      4      6
3      3      9      2
4      4      2      3
5      5      9      1
6      6      6      7
> #Defino dos vectores
> summer = (satisfaction$summer)
> winter = (satisfaction$winter)
> summer
 [1]  7  4  9  2  9  6  9  1  4  1  3 10  5  9  1 10  5  9  8  9  6 10  1  9  2
[26]  3 10  5  4  1
> winter
 [1] 9 6 2 3 1 7 5 2 8 5 6 4 1 9 5 9 6 2 4 1 7 5 2 8 5 6 4 5 8 4
> #Calculo de medias de cada vector
> msummer = mean( summer )
> mwinter = mean( winter )
> msummer
[1] 5.733333
> mwinter
[1] 4.966667
> #Calculo de varianza de cada vector
> vsummer = var( summer )
> vwinter = var ( winter )
> vsummer
[1] 11.02989
> vwinter
[1] 6.171264
> VV = vsummer / vwinter

> VV

[1] 1.787297

> #Calculo de desviación standart de cada vector
> desvsummer = sd( summer )
> desvwinter = sd ( winter )
> desvsummer
[1] 3.321127
> desvwinter
[1] 2.484203
> #Longitud del fichero
> n = length(satisfaction[[1]])
> n
[1] 30
> #Longitud del cada uno de los vectores
> nsummer = length(summer)

> nwinter = length(winter)

> nsummer

[1] 30

> nwinter

[1] 30
```

Nos planteamos una diferencia de medias:

La Hipótesis nula será la que corresponde a que la satisfacción media de un cliente que ha viajado tanto en invierno como en verano es igual.

La Hipótesis alternativa será que la satisfacción media dé un cliente que ha viajado tanto en invierno como en verano es diferente.

H0 : μsummer = 1 μwinter y

H1: μsummer ≠ 1 μwinter

α = 5%. => 0.05

#Intervalo de confianza para la media summer con 95%

> msummer + desvsummer * qt(0.25, n-1)/sqrt(n)

[1] 5.319168

> msummer - desvsummer * qt(0.25, n-1)/sqrt(n)

[1] 6.147498

#Intervalo de confianza para la media winter con 95%

> mwinter + desvwinter * qt(0.25, n-1)/sqrt(n)

[1] 4.656871

> mwinter - desvwinter * qt(0.25, n-1)/sqrt(n)

[1] 5.276462

Varianza Residual

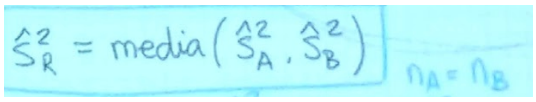
ResiVar = mean (c ((vsummer), (vwinter)))

> ResiVar

[1] 8.600575

>

Sabiendo que la varianza residual es igual a la media de la varianza de summer y la media de la varianza de winter, cuando cuando el número de observaciones del primer tratamiento y el numero de observaciones del segundo tratamiento son iguales => nsummer = nwinter.



> numerador = msummer - mwinter

> denominador = sqrt(ResiVar) * sqrt (1/nsummer + 1/nwinter)

> t0 = numerador / denominador

> t0

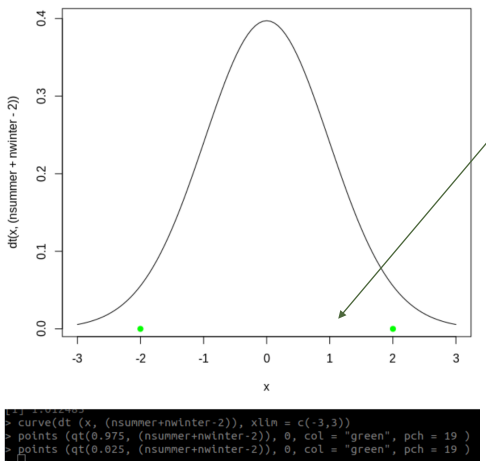
[1] 1.012485

$$t_0 = \frac{\bar{y}_{1\bullet} - \bar{y}_{2\bullet}}{\hat{s}_R \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \rightarrow t_{n-2}$$

> curve(dt (x, (nsummer+nwinter-2)), xlim = c(-3,3))

> points (qt(0.975, (nsummer+nwinter-2)), 0, col = "green", pch = 19)

> points (qt(0.025, (nsummer+nwinter-2)), 0, col = "green", pch = 19)



> t0

[1] 1.012485

Al calcular el t0 y ubicarlo en la curva podemos ver que esta entre los límites

```
> wilcox.test(summer,winter)

Wilcoxon rank sum test with continuity correction

data:  summer and winter
W = 515.5, p-value = 0.333 > (α=0.05) => acepto H0

alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(summer, winter) :
  cannot compute exact p-value with ties
```

Las dos muestras

Contraste de la wilcox test

- p-value = seguridad para aceptar ó rechazar Hipótesis

```
> t.test(winter, summer, alternative = "two.sided", var.equal=T,
con.level=0.95)

Two Sample t-test

data:  winter and summer
t = -1.0125, df = 58, p-value = 0.3155 > (α=0.05) => acepto H0

alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.282393  0.749060
sample estimates:
mean of x mean of y
 4.966667  5.733333
```

Las dos muestras

Contraste de la t-student

- t = donde está posicionado el estadístico
- df = grados de libertad
- p-value = seguridad para aceptar ó rechazar Hipótesis

hypothesis alternativa: indica que es bilateral

Intervalo de confianza: 95 % que -> la diferencia de las muestras esta entre

-2.282393 0.749060

sample estimates: medias estimadas

mean of x mean of y

4.966667 5.733333

```
> var.test(summer, winter)

F test to compare two variances

data:  summer and winter
F = 1.7873, num df = 29, denom df = 29, p-value = 0.1237 > (α=0.05) => acepto H0
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.8506906 3.7551045
sample estimates:
ratio of variances
         1.787297

>
```

Las dos muestras

Contraste de la var.test

- df = grados de libertad
- p-value = seguridad para aceptar ó rechazar Hipótesis

hypothesis alternativa: indica que es bilateral

Intervalo de confianza: 95 % que -> la diferencia de las muestras esta entre

0.8506906 3.7551045

sample estimates: ratio of variances

1.787297

Los cálculos realizados anteriormente tanto para la media como para la varianza nos muestran que no podemos estar de acuerdo con el CEO de Avio, ya que no existen evidencias suficientes para aceptar la Hipótesis alternativa H1.

5. Avio has been asked to give a quote for a group. You can either offer a full fare price of 500€ or a discount fare price of 350€. Based on past experience the probability of the full fare price to be accepted is 65%. Instead, if you offer the discount fare price you are almost certain they will buy the tickets (100% probability).

Open the given Excel file (prescriptive_group.xlsx), fill the decision tree template with the given figures, and make the necessary calculations.

Should the company offer the discount or full fare price?

What if the probability of selling the full fare ticket is 70%?

Avio –> presupuesto para un grupo.

A: Tarifa completa de 500€
Casos favorables = 65%
Casos posibles = 100%

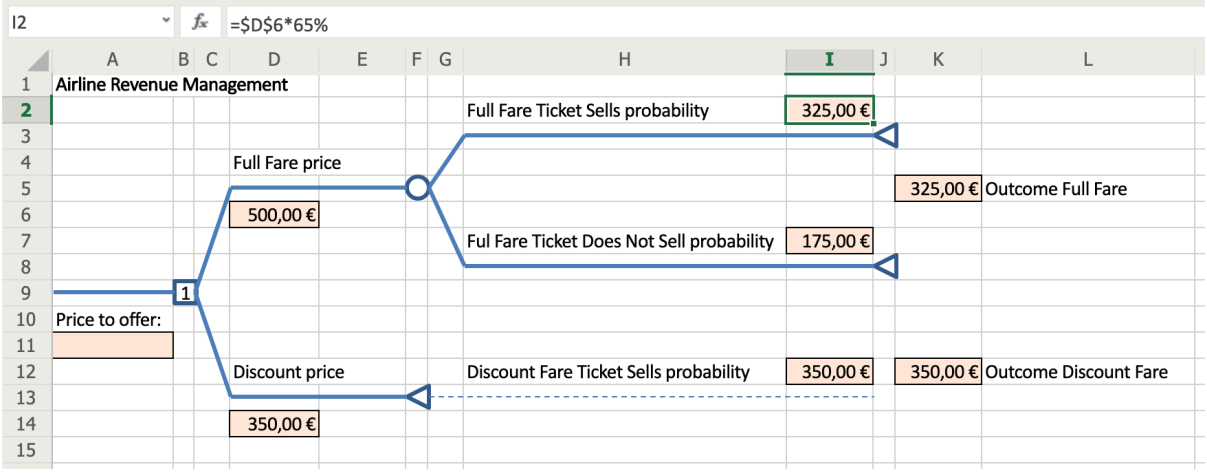
0.65 + 0.35 = 1 => 325€ + 0 = 325€

B. tarifa de descuento de 350€
Casos favorables = Casos posibles = probabilidad 100%

1 => 350€

De acuerdo con los resultados obtenidos:

¿La empresa debe ofrecer => tarifa completa ya que es mucho más alto el beneficio para la empresa.

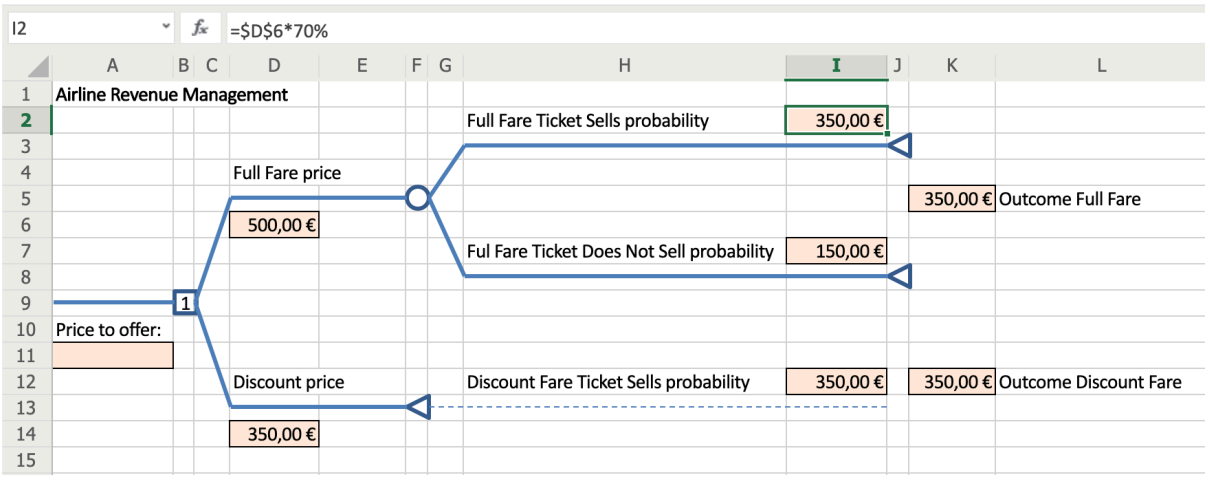


A: Tarifa completa de 500€
Casos favorables = 70%
Casos posibles = 100%

0.70 + 0.30 = 1 => 350€ + 0 = 350€ €

B. tarifa de descuento de 350€
Casos favorables = Casos posibles = probabilidad 100%

1 => 350€



De acuerdo con los resultados obtenidos:

Si la probabilidad de vender el billete de tarifa completa es del 70%, cualquiera de las dos opciones es buena ya que el beneficio obtenido por la empresa es el mayor.