Pronosticar el tiempo de vida de unidades de almacenamiento.

Contexto

Cada día, Backblaze toma una foto de cada disco duro operativo que incluye información básica del disco duro (por ejemplo, capacidad, falla) y S.M.A.R.T. estadísticas reportadas por cada unidad. Este conjunto de datos contiene datos de los dos primeros trimestres de 2016.

Contenido

Este conjunto de datos contiene información básica del disco duro y 90 columnas o valores sin procesar y normalizados de 45 S.M.A.R.T. estadísticas diferentes. Cada fila representa una observación diaria de un disco duro.

- date: Fecha en formato yyyy-mm-dd
- serial_number: Número de serie del disco asignado por el fabricante.
- model: Número de modelo del disco asignado por el fabricante.
- capacity_bytes: Capacidad del disco en bytes.
- failure: Contiene un "0" si el disco está OK. Contiene un "1" si el último día la unidad estuvo operativa antes de fallar.
- 90 variables que comienzan con smart': valores sin procesar y normalizados para 45 estadísticas SMART diferentes según lo informado para una unidad dada

Hints

Algunos elementos a tener en cuenta al procesar los datos:

- Las estadísticas SMART pueden variar en significado según el fabricante y el modelo. Puede ser más informativo comparar unidades que sean similares en modelo y fabricante.
- Algunas columnas SMART pueden tener valores fuera de límites
- Cuando falla una unidad, la columna 'falla' se establece en 1 el día de la falla y, a partir del día siguiente, se eliminará la unidad del conjunto de datos. Cada día, también se agregan nuevas unidades. Esto significa que la cantidad total de unidades por día puede variar.
- SMART 9 es la cantidad de horas que una unidad ha estado en servicio. Para calcular la edad de una unidad en días, divida este número entre 24.

1. Parte

Limpieza de datos e ingeniería de variables

a) Transforme los datos a tiempos de vida de unidades.

Solución:
Lo primero realizado fue eliminar todas las columnas raw a excepción de la columna smart 9 raw debido a que esta es el tiempo de vida de los discos. A continuación, se dividió en 24 la columna smart 9 raw, para tener el tiempo de vida en días.

Code:

```
columnas=df.columns
columnas_raw=[]
for i in columnas:
    l=len(i)
    if i[l-3:]== 'raw':
        columnas_raw.append(i)
columnas_raw.append('smart_9_normalized')
# no eliminamos la columna smart_9_raw,
# ya que esta indica el tiempo de vida del disco
columnas_raw.remove('smart_9_raw')
df=df.rename(columns={'smart_9_raw':'time_day'})
df['time_day']=df['time_day']/24
df = df.drop(columns=columnas_raw, axis=1)
```

Para realizar este proceso, se descartaron todos los items que no fallaron, ya que el enfoque es identificar fallas en los dispositivos. Esto genera un sesgo ya que los discos que se mantienen en buen estado durante el estudio no son contemplados.

Para solo guardar la información de los discos que alguna vez fallaron se generó un vector con todos los números de serie de los items que fallaron y guardaron en una base de datos todos estos items.

Code:

```
failed_hdds = df.loc[df.failure==1]["serial_number"]
failed_hdds.describe()
```

Según los visto con la función `.describe()` de estos Discos que alguna vez fallaron se encontraron que hay 10 registros duplicados de fallas, los cuales se procedió a eliminar, llegando así a 205 Discos a analizar en este estudio.

Code:

```
df = df.loc[df["serial_number"].isin(failed_hdds)]
df.to_csv("data_fail.csv", index = False, sep=',', encoding='utf-8')
```

Ahora podemos trabajar con información no censurada.

Para trabajar con los tiempos de vida de las unidades, se tomó en cuenta la variable SMART 9, al momento de fallar de los distintos dispositivos. de esta forma, se obtuvieron los tiempos de vida de

los dispositivos al momento de fallar, pero en horas, por lo que se procedió a dividir estos valores por 24, para obtener el tiempo de vida en días y trabajar con una escala inferior.

b) Calcule features de las variables SMART correspondientes a promedio, desviaciones estándar y coeficiente de asimetría, de la última semana, últimas dos semanas, último mes.

Solución:

1 En esta parte, se apreció que existen variables smart que no variaron durante el tiempo, ni entre los items.

Por lo que se decidió descartar todas las variables smart que posee una covarianza inferior a 0,1.

2 Se verificó cuántos "Nas"hay de cada variables SMART, que quedaron de la eliminación. De lo que se obtuvo el siguiente gráfico:

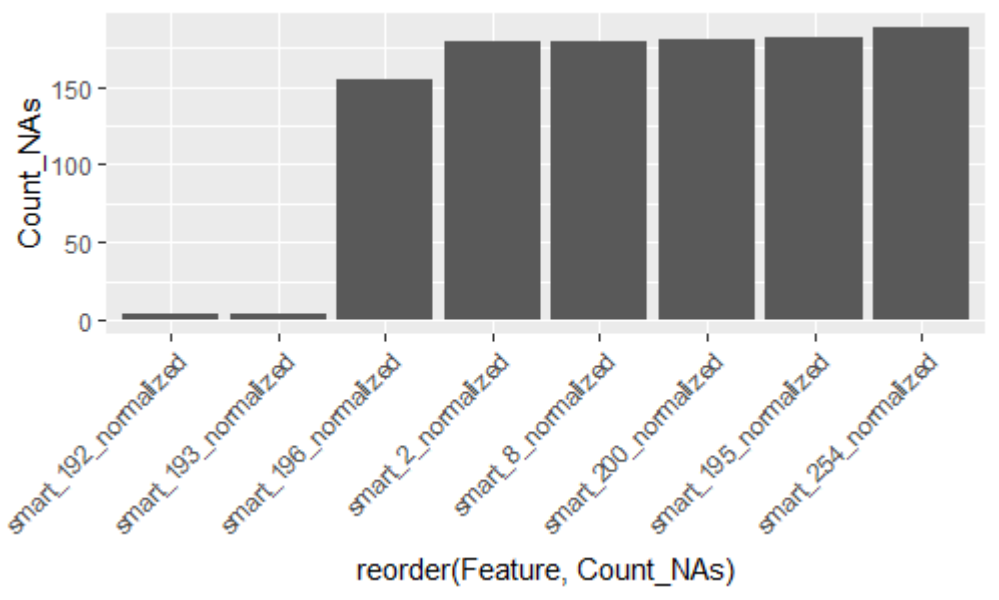

Figura 1: Histograma de Nas

3 De estas variables SMART se conservaron

- smart_192_normalized
- smart_193_normalized

Las cuales solo poseían 4 items que no estaban registrados.

4 Se verificó además $\frac{\mu}{\sigma}$ de las otras variables, como la capacidad. Se aprecia que esta es de 0, por lo que todos los discos que fallaron eran de la misma capacidad, por lo que no se considera un factor importante en la identificación de comportamiento de fallas.

5 De la tabla de media, covarianza, desviacion, kurtosis(skew), se pasó a una gráfica para apreciar las variaciones temporales, de los valores de las variables smart.

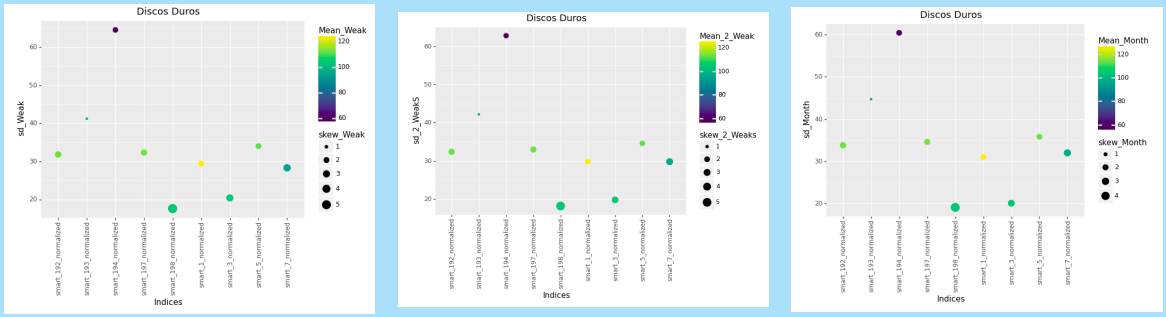

Figura 2: Metricas al detalle.

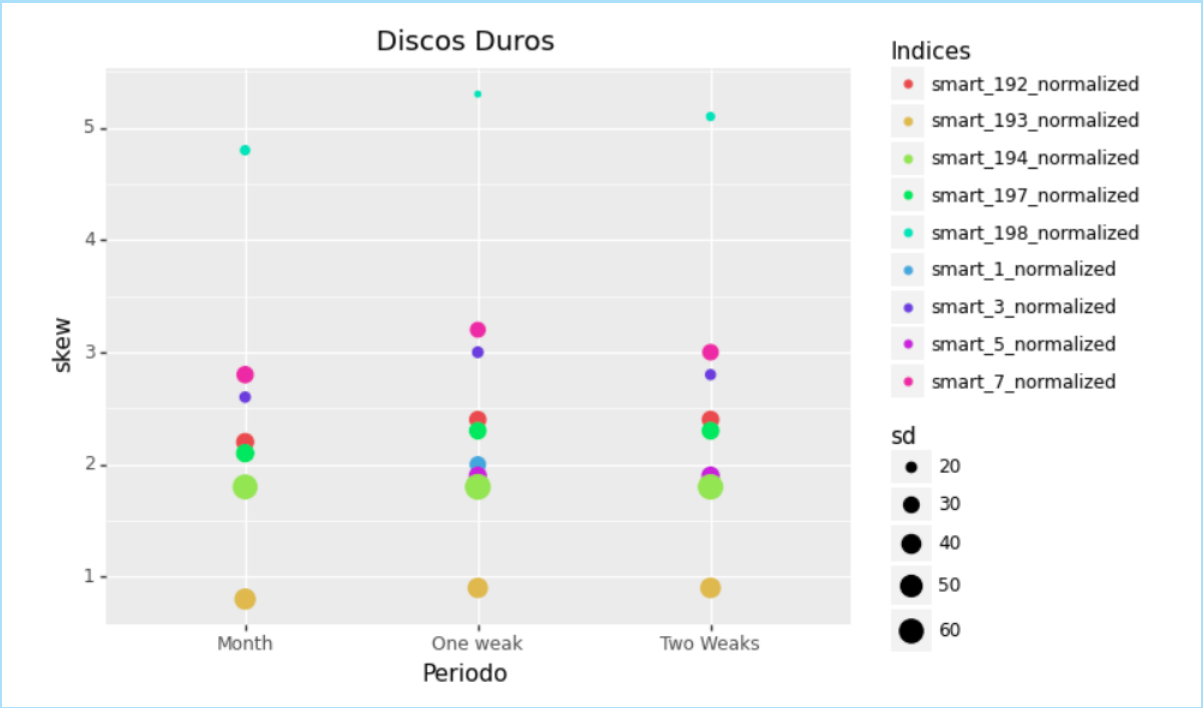

Figura 3: Métricas por periodo de tiempo

Se aprecia que estas variaciones no son significativas en el tiempo, por lo que se puede ver que no incidían claramente en un deterioro significativo en todos los discos duros.

Por un lado, pese a que no se aprecien variaciones temporales significativas en las distribuciones de las variables smart, no se descartan que los pequeños cambios coincidan en cambios en el tiempo de vida de los productos. Es por ello que se generaron 9 variables que proceden de el cambio que sufre el Skew, de la última semana con la penúltima semana(Skew[Última semana]- Skew[Penúltima semana]).

Por otro lado, se consideraron las nueve variables smart del análisis, para los siguientes pasos, en donde son representadas como el promedio de los valores que se tuvieron para cada disco.

2. Parte

Análisis descriptivo de los tiempos de sobrevivencia.

a) Realice un análisis descriptivo bivariado entre tiempo de vida de las unidades a través de las curvas de Kaplan-Meier con cada una de las features.

b) Realice Test de comparación de curvas

- **Responderemos a) y b) a continuación**

Solución:
Respuesta de a) y b):

lo primero a realizar fue una gráfica de Kaplan-Meier para visualizar el $S(t)$ general. para todas las unidades.

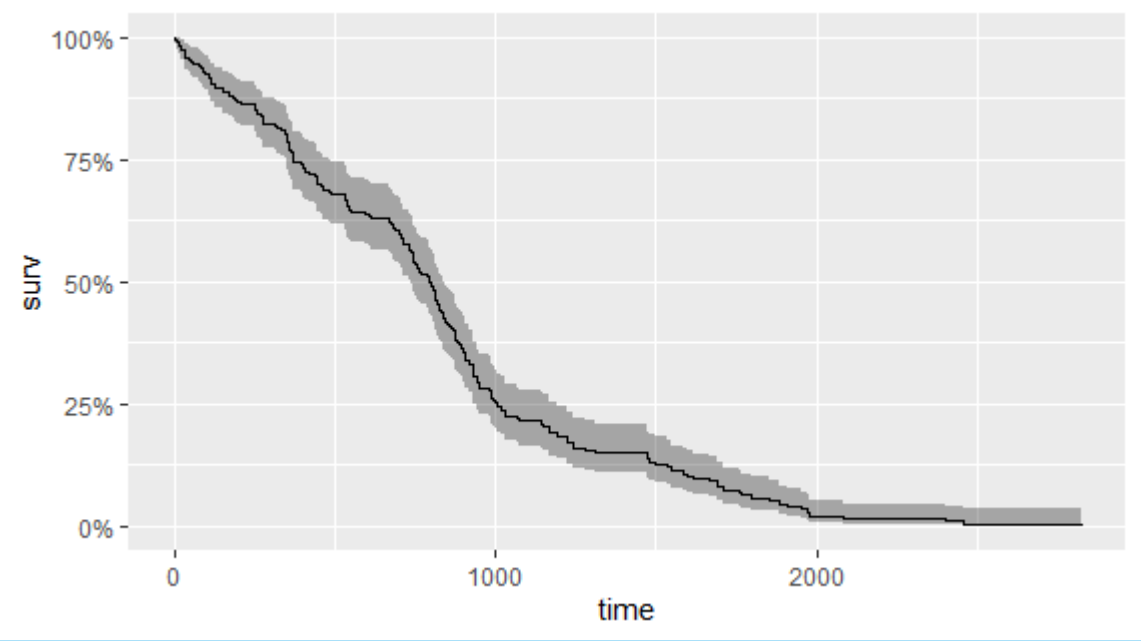

Figura 4: KM General

Se aprecia que existen unidades que fallaron en pocos días y hubo unas pocas que fallaron ya después de 2000 días de funcionamiento.

A continuación se realizará las gráficas de KM para cada una de las variables que nos quedamos, las cuales son 19. Las cuales debemos volver categóricas, de a lo mas 4 dimensiones, lo cual permite a la función de KM graficar una cantidad de curvas que permite ser interpretadas y representativas en cierta forma. Luego, se debe verificar a partir de su comportamiento en los gráficos de kaplan-meier, si deben mantenerse como variables categóricas o convertirse a numéricas.

Variable Model:

Para segregar los modelos se visualizó su frecuencia y distribución en los datos. Para identificar posibles agrupaciones.

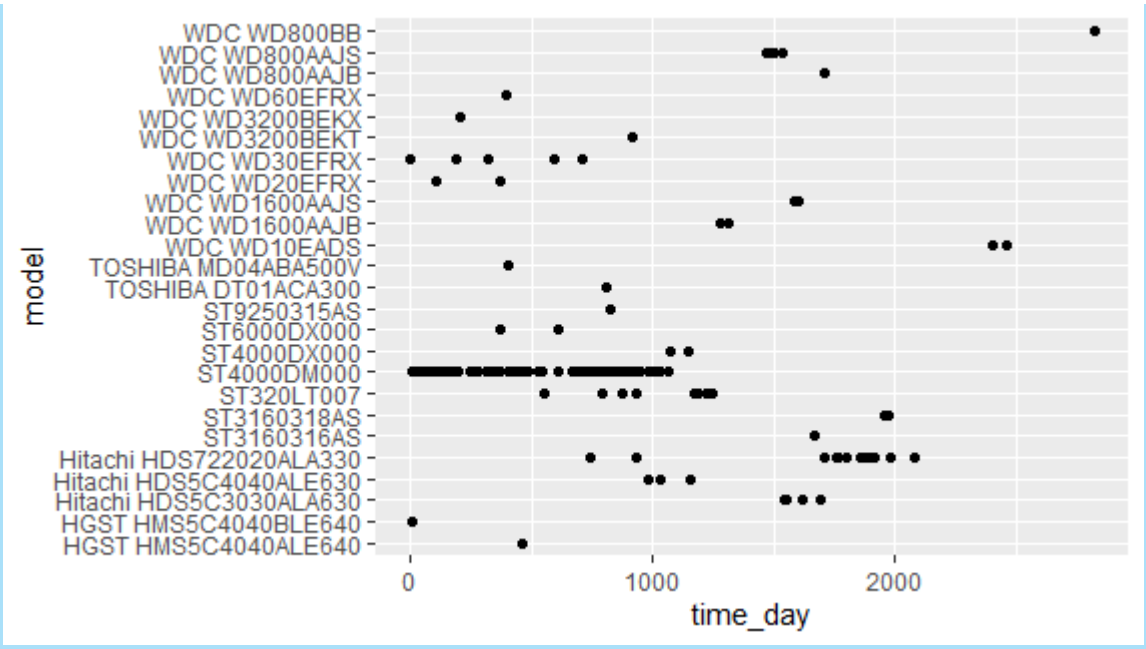

Figura 5: Gráfica de puntos para modelo

Se aprecia que hay un modelo predominante, que posee una dispersión de los tiempos de vida bastante acotada.

Ahora veremos cuantos discos en cada modelo existen.

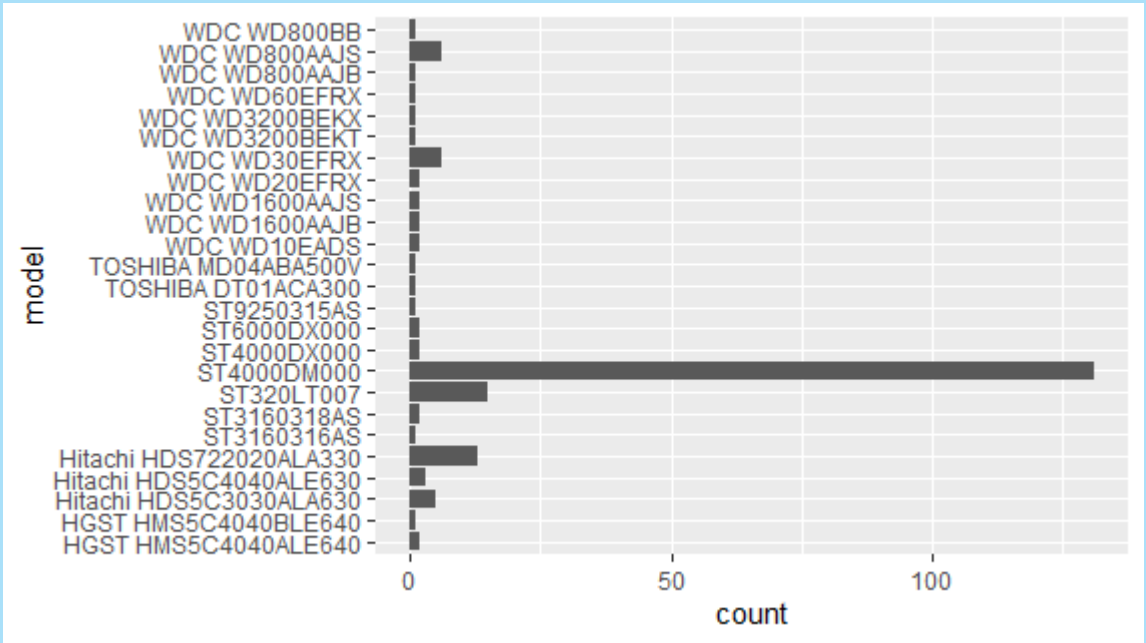

Figura 6: Histograma de modelos

En base a la distribución se decidió categorizar los discos, como modelo ST4000DM00 y other.

Según esta categorización el resultado de las curvas de km es el siguiente:

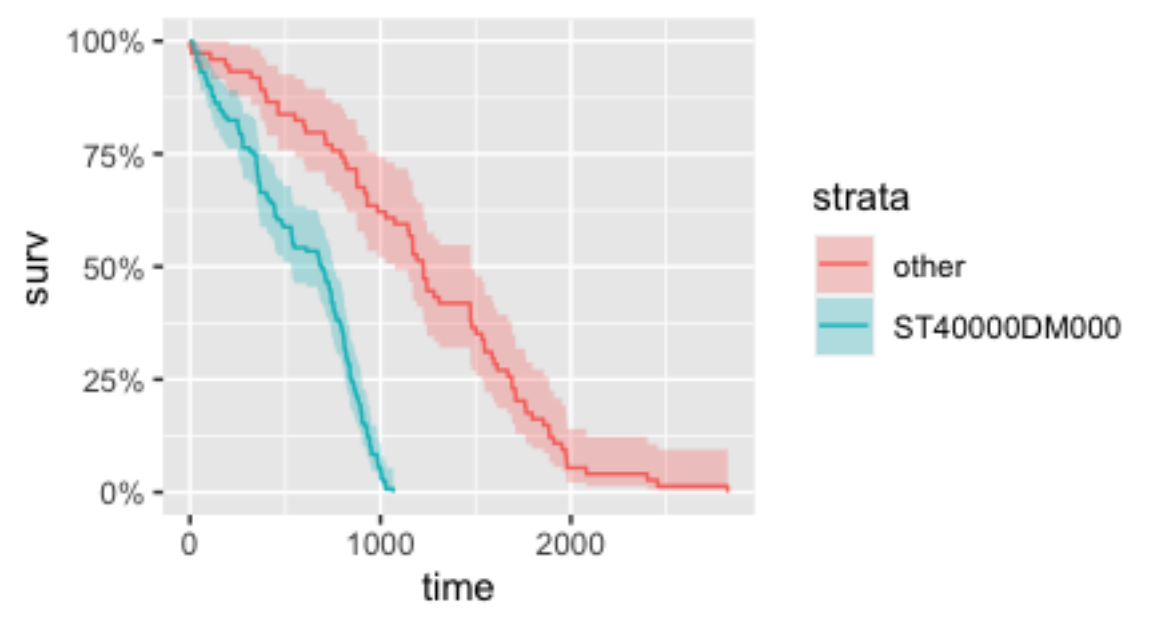

Figura 7: Histograma de modelos

Según la gráfica anterior, se aprecia que las curvas son distintas entre sí, para asegúranos se realizan dos test de comparación de curvas, en donde se entrega el valor P de cada uno, si este es inferior a 0,005 significa que hay evidencia para rechazas H_0, hipótesis que indica que las curvas presentadas en la gráfica 7 son idénticas.

Cuadro 1: Test de comparación de curvas, para model

Test	**Valor P**	**Conclusión**
LogRank	2.233694e-20	*Hay evidencia suficiente como para rechazar* H_0
Peto	6.057151e-13	*Hay evidencia suficiente como para rechazar* H_0

Los test realizados son LogRank y el test de Peto, el primero realiza una comparación que le asigna mas peso a la cola de la derecha de la curva, es decir a los tiempos de vida mas altos y menos importancia a las diferencias que existen con las muertes mas tempranas. El segundo test, le asigna una mayor importancia a las muertes prematuras, por lo tanto nos indica un punto de vista distinto del test LogRank. Tomando en cuenta ambos test, podemos asegurarnos que las curvas en el gráfico 7 sean distintas en el comienzo como en el final.

Dado el gráfico anterior(7) y el cuadro de los test de comparación de curvas 1 se aprecia que hay evidencia para rechazar H_0. Por lo tanto para los pasos siguientes esta variable se considerará como categórica, manteniendo la categorización realizada para la gráfica 7.

Variables smart

Para las variables smart se siguieron los siguientes pasos para la categorización de sus valores o agrupación.

1. Se verificó su distribución a través de un histograma, con el cual se tomó una decisión de como particionar los discos en base a la variable.
2. Se configuró una división en base a un Rango de valores, otorgados por una división en percentiles.
   Por lo que, si la variable se podía dividir en cuartiles el resultado son 4 categorías para dicha variable. En cambio hay variables smart que poseen un comportamiento que no permite esta división y se debe ver de qué forma dividir los datos haciendo uso de la misma estructura.
   Por ejemplo, 1 es la categoría para los discos que poseen un valor smart que esta en el Rango del 75 % y 100 % de los datos.
3. Con las categorías definidas, se categorizó cada disco y se construyeron las curvas de Kaplan meier.

Variables Smart Normalizadas

Smart 1 Normalized

Se separó en cuartiles.

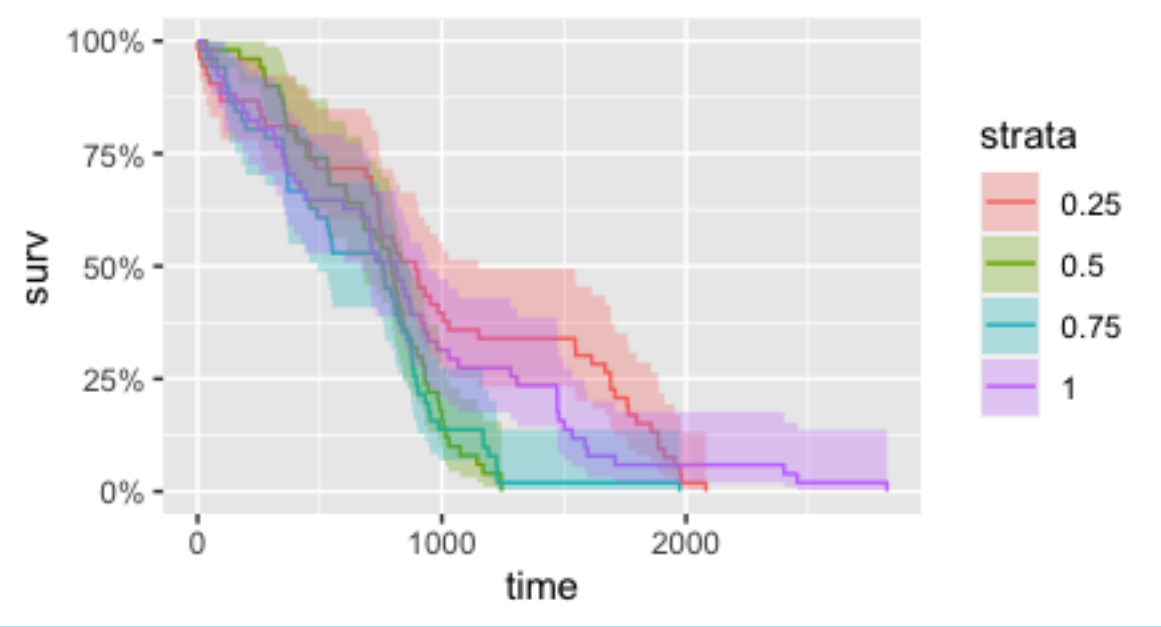

Figura 8: Kaplan Meier Smart 1

Como se aprecia en este gráfico, no se presenta linealidad en la categorización, por lo que volver la variable a una numérica carece de sentido. Lo que significa que esta variable permanecerá como una categórica.

Cuadro 2: Test de comparación de curvas, para smart_1_normalized

Test	Valor P	Conclusión
LogRank	0.001013532	*Hay evidencia suficiente como para rechazar* H_0
Peto	0.1024773	*Se conserva* H_0*, No hay evidencia para rechazar la hipótesis nula*

A partir del gráfico 8, se puede observar que existe un constante cruce entre las curvas, lo que nos puede guiar a que estas curvas no sean representativas. Además, respecto al cuadro 2 y al test de Peto, se tiene evidencia suficiente para conservar H_0, lo que significa que la hipótesis nula de que las curvas son distintas, se cumple. Es por ello que, se descartará esta variable para los análisis posteriores, debido a que esta difícilmente logra representar el tiempo de vida de los discos duros.

Smart 3 Normalized

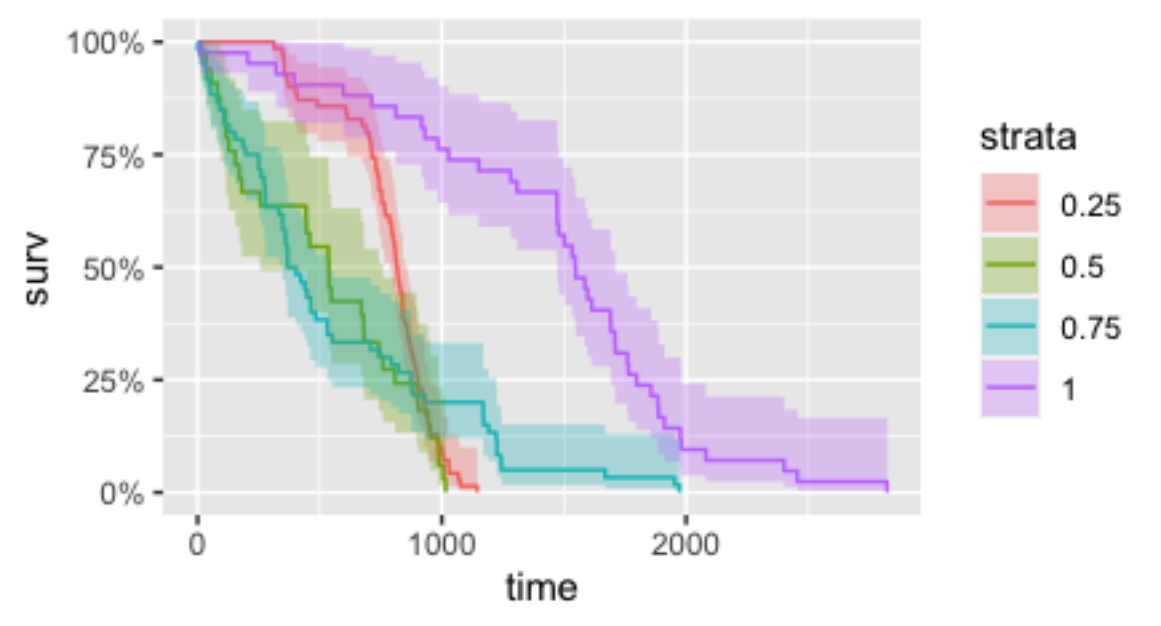

Figura 9: Kaplan Meier Smart 3

Al igual que con el gráfico anterior (8), no se presenta una linealidad entre las categorías de las variables, por lo que nuevamente se mantendrá a la variable como una categórica.

Cuadro 3: Test de comparación de curvas, para smart_3_normalized

Test	Valor P	Conclusión
LogRank	1.051473e-15	*Hay evidencia suficiente como para rechazar* H_0
Peto	6.331266e-14	*Hay evidencia suficiente como para rechazar* H_0

Dado el gráfico 9 y el cuadro 3 se aprecia que hay evidencia para rechazar H_0 en ambos test. Por lo

que se procede de la misma forma que con la variable anterior.

Smart 5 Normalized

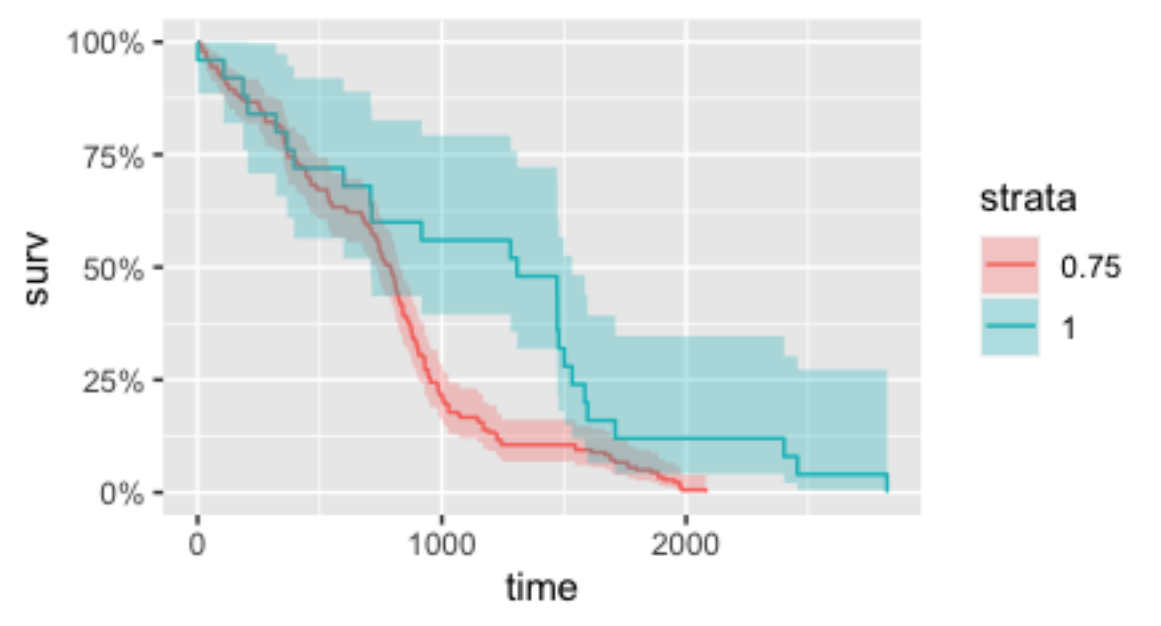

Figura 10: Kaplan Meier Smart 5

A diferencia del gráfico anterior (9), en este si se puede apreciar una linealidad entre las categorías, por lo que se decidió no mantener la variable como categórica y convertirla a numérica.

Cuadro 4: Test de comparación de curvas, para smart_5_normalized

Test	**Valor P**	**Conclusión**
LogRank	0.002104752	*Hay evidencia suficiente como para rechazar* H_0
Peto	0.03913104	*Hay evidencia suficiente como para rechazar* H_0

Dado el gráfico 10 y el cuadro 4 se aprecia que hay evidencia para rechazar H_0 en ambos test. Por lo que se procede de la misma forma que con la variable anterior.

Smart 7 Normalized

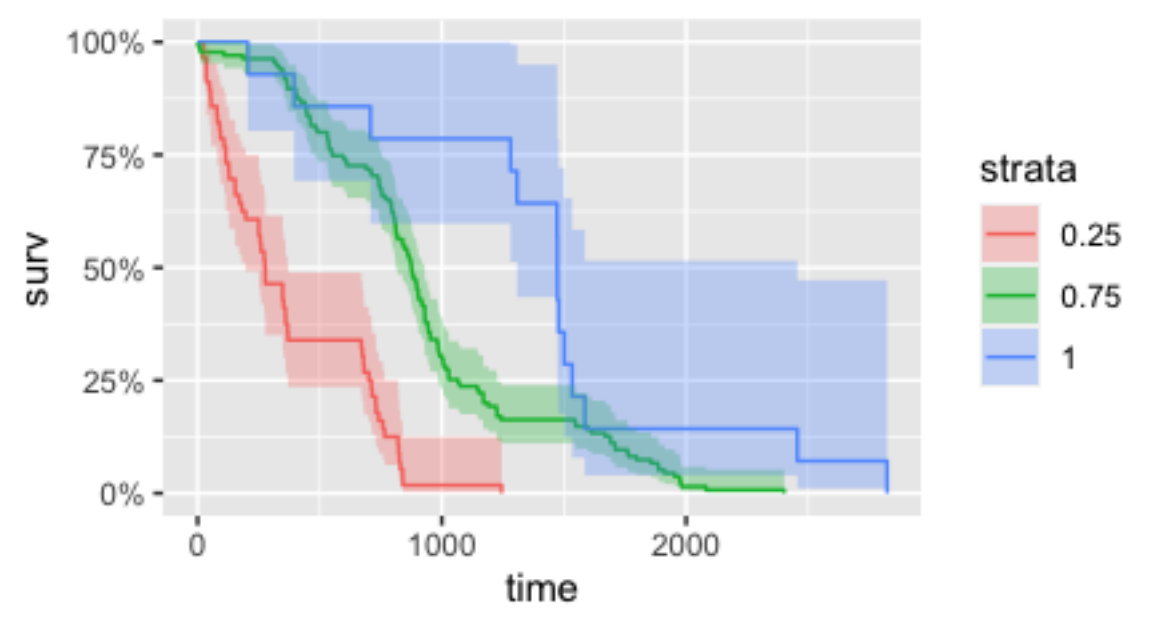

Figura 11: Kaplan Meier Smart 7

Del mismo modo a lo que ocurrió en el gráfico 10, se convertirá a esta variable a una numérica, debido a que se observa una linealidad entre las categorías de la variable.

Cuadro 5: Test de comparación de curvas, para smart_7_normalized

Test	**Valor P**	**Conclusión**
LogRank	1.041497e-22	*Hay evidencia suficiente como para rechazar* H_0
Peto	6.650813e-22	*Hay evidencia suficiente como para rechazar* H_0

Dado el gráfico 11 y el cuadro 5 se aprecia que hay evidencia para rechazar H_0 en ambos test. Por lo que se procede de la misma forma que con la variable anterior, entonces se utilizará en los pasos siguientes como numero y no como formato categórico.

Smart 192 Normalized

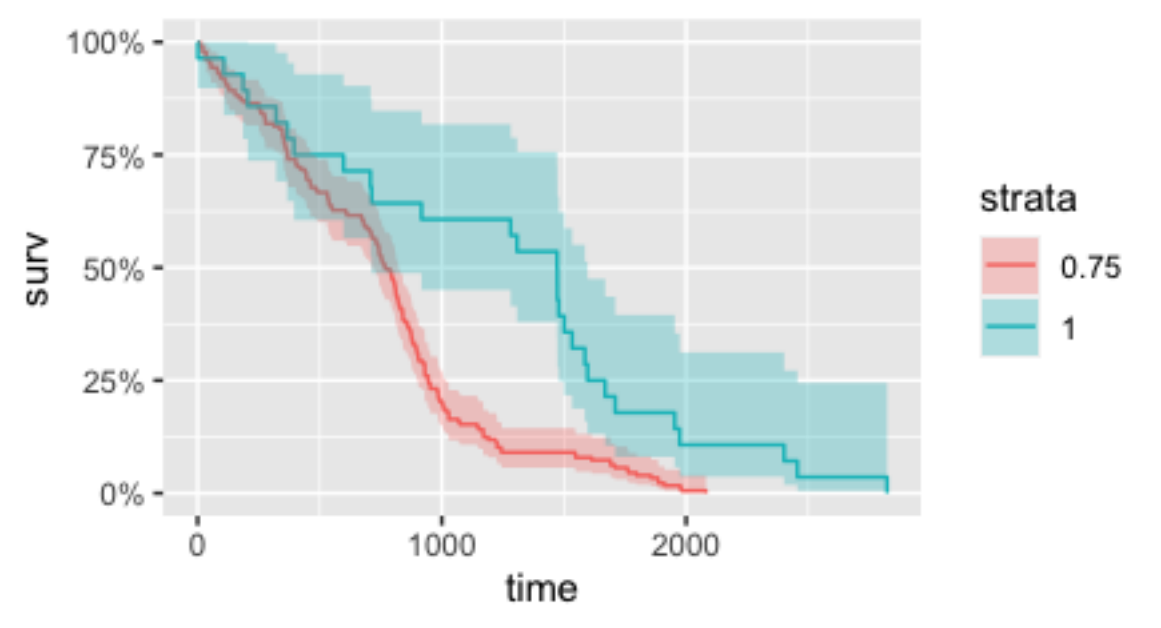

Figura 12: Kaplan Meier Smart 192

Al igual que con los gráficos 10 y 11, la relación entre las categorías es lineal, por lo que se puede realizar una conversión de la variable, de una categórica a una numérica.

Cuadro 6: Test de comparación de curvas, para smart_192_normalized

Test	**Valor P**	**Conclusión**
LogRank	8.055711e-05	*Hay evidencia suficiente como para rechazar* H_0
Peto	0.004568904	*Hay evidencia suficiente como para rechazar* H_0

Dado el gráfico 12 y el cuadro 6 se aprecia que hay evidencia para rechazar H_0 en ambos test. Por lo que se procede de la misma forma que con la variable anterior, entonces se utilizará en los pasos siguientes como numero y no como formato categórico.

Smart 193 Normalized

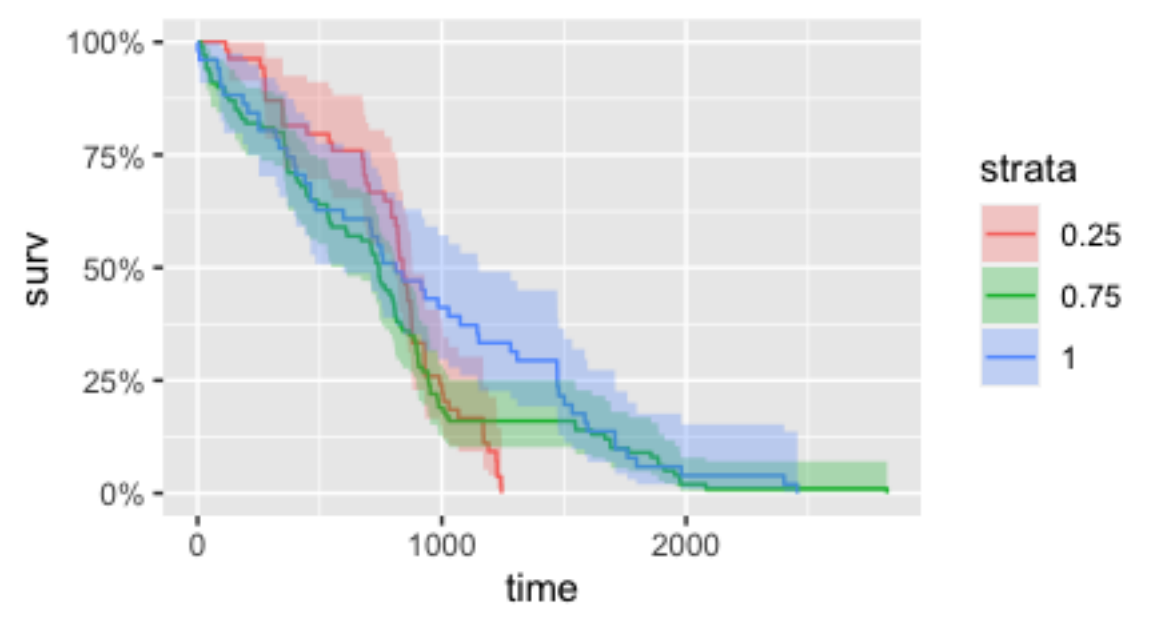

Figura 13: Kaplan Meier Smart 193

En este gráfico, no se observa una linealidad, por lo que la variable se mantendrá como categórica.

Cuadro 7: Test de comparación de curvas, para smart_193_normalized

Test	**Valor P**	**Conclusión**
LogRank	0.1977068	*Se conserva H_0, No hay evidencia para rechazar la hipótesis nula*
Peto	0.2680677	*Se conserva H_0, No hay evidencia para rechazar la hipótesis nula*

Dado el gráfico 13 se aprecia que las curvas se cruzan constantemente por lo que puede que estas curvas no sean representativas y en el cuadro 7 se aprecia que no hay evidencia para rechazar H_0 en ambos test, esto quiere decir que no hay evidencia para señalar que las curvas son distintas. Tomando en cuenta que no se aprecia una correlación clara entre esta variable y la variable tiempo de supervivencia se descarta para utilizarla en los modelos.

Smart 194 Normalized

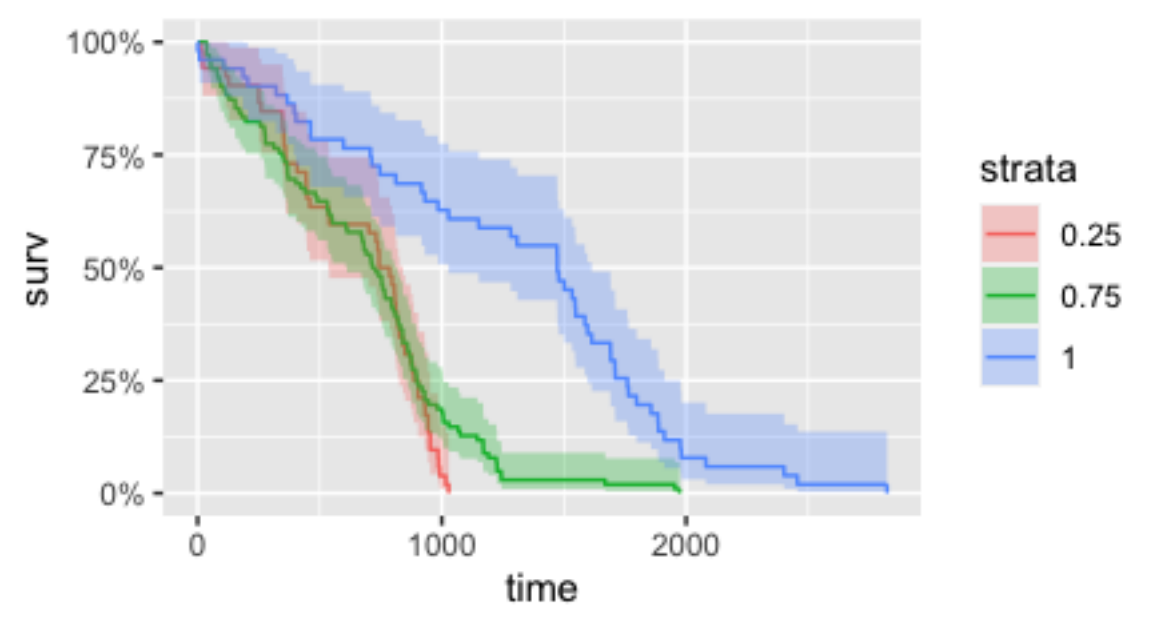

Figura 14: Kaplan Meier Smart 194

Del mismo modo a lo ocurrido en los casos 9, 10, 11 y 12, se convertirá a la variable en una numérica, debido a que se presenta una linealidad entre las categorías de la variable.

Cuadro 8: Test de comparación de curvas, para smart_194_normalized

Test	**Valor P**	**Conclusión**
LogRank	1.014837e-11	*Hay evidencia suficiente como para rechazar* H_0
Peto	1.438931e-06	*Hay evidencia suficiente como para rechazar* H_0

Dado el gráfico 14 y el cuadro 8 se aprecia que hay evidencia para rechazar H_0 en ambos test. Por lo que se procede de la misma forma que con la variable anterior, entonces se utilizará en los pasos siguientes como numero y no como formato categórico.

Smart 197 Normalized

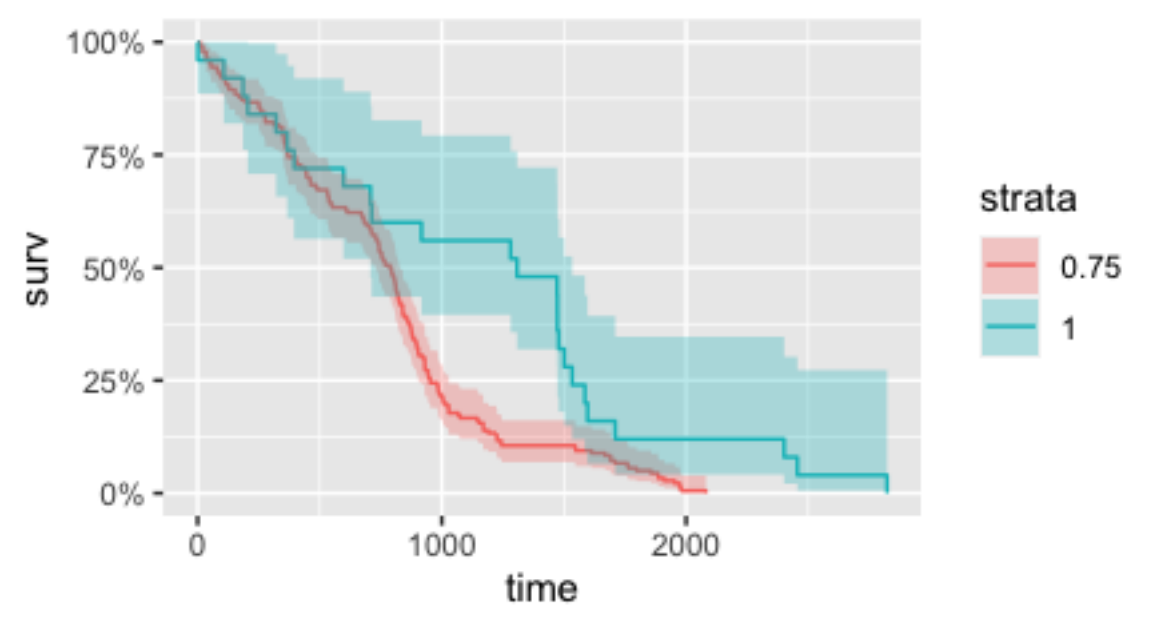

Figura 15: Kaplan Meier Smart 197

Del gráfico anterior se desprende que, existe linealidad entre las categorías, por lo que se puede convertir esta en una numérica. Al igual que lo que ocurrió en los casos 9, 10, 11, 12 y 14.

Cuadro 9: Test de comparación de curvas, para smart_197_normalized

Test	**Valor P**	**Conclusión**
LogRank	0.002104752	*Hay evidencia suficiente como para rechazar* H_0
Peto	0.03913104	*Hay evidencia suficiente como para rechazar* H_0

Dado el gráfico 15 y el cuadro 9 se aprecia que hay evidencia para rechazar H_0 en ambos test. Por lo que se procede de la misma forma que con la variable anterior, entonces se utilizará en los pasos siguientes como numero y no como formato categórico.

Smart 198 Normalized

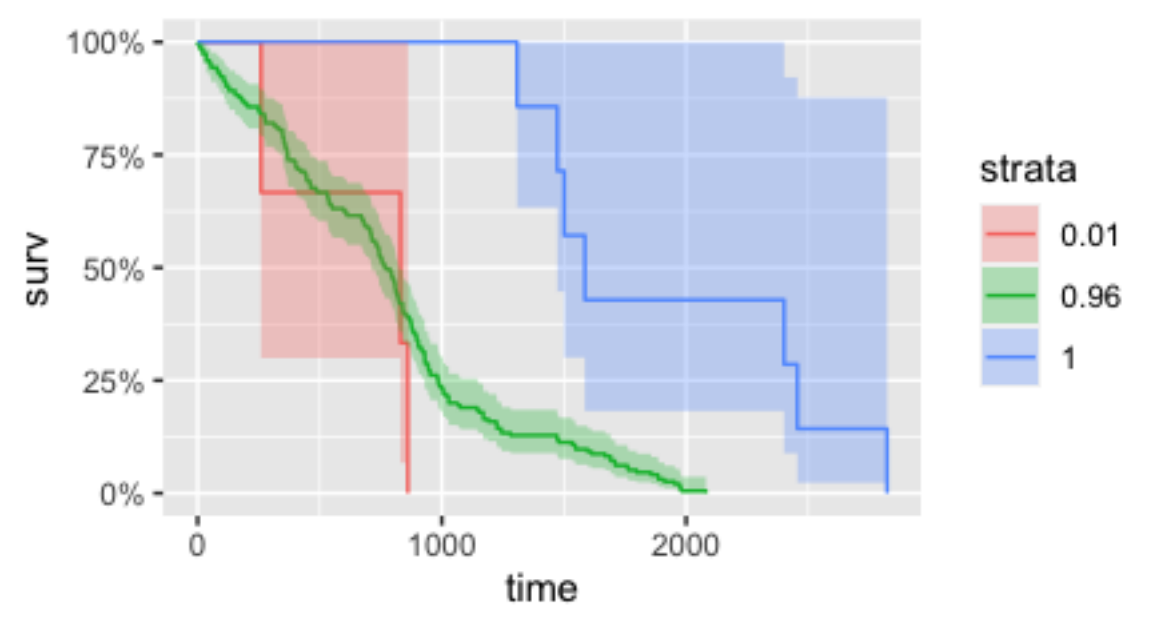

Figura 16: Kaplan Meier Smart 198

En este caso en particular, a pesar que para las categorías 0.01 y 0.96 se observan cruces que abarcan un mayor grado y con lo cual se tendería a dejar a la variable como categórica, esta se convertirá a numéricas, debido a que como se observa del gráfico, la categoría 0.01 tiene aproximadamente 3 registros y con ello no se justifica que la variable se mantenga como categórica por falta de linealidad de sus categorías.

Cuadro 10: Test de comparación de curvas, para smart_198_normalized

Test	**Valor P**	**Conclusión**
LogRank	0.000249199	*Hay evidencia suficiente como para rechazar* H_0
Peto	0.00476285	*Hay evidencia suficiente como para rechazar* H_0

Pese a los pocos datos utilizados en la categorización el resultado del gráfico 14 y el cuadro 8 representan que hay evidencia para rechazar H_0 en ambos test. Por lo que se procede de la misma forma que con la variable anterior, entonces se utilizará en los pasos siguientes como numero y no como formato categórico.

Variables smart skwes

Para realizar el análisis bivariado de Kaplan Meier(KM) se tomará en cuenta las diferencias de los skews dentro de las variables smart normalized. Estas diferencias son valores numéricos que atraves de KM no pueden ser gráficadas. por lo cual se procedió de igual forma que con las variables smart, categorizando por medio de divisiones de los valores de cada columna. Cuando se encontraban etiquetas

que en Kaplan Meier se comportaban de igual forma y podian ser agrupadas por medio de intervalos numéricos, por ejemplo que los primeros 3 cuartiles posean curvas muy similares, estos se etiquetaban como una sola categoría. Tomando en cuenta la categorización a continuación se encuentran los gráficos KM resultantes de estas categorías.

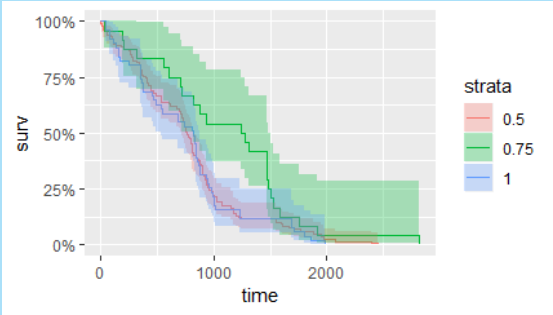

Figura 17: km Smart 1 Normalized Delta Skews

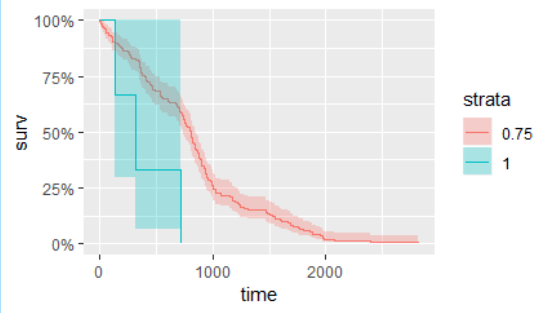

Figura 18: km Smart 3 Normalized Delta Skews

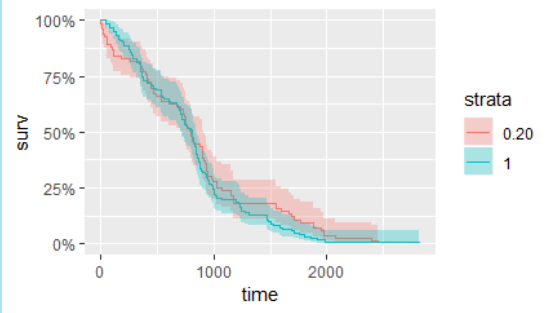

Figura 19: km Smart 5 Normalized Delta Skews

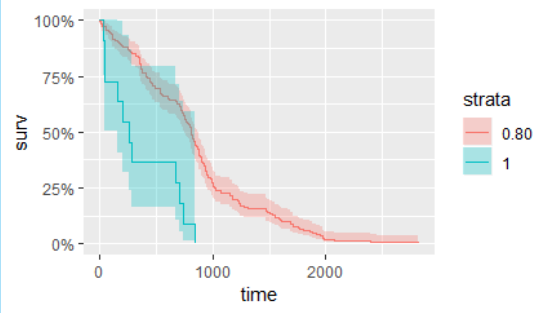

Figura 20: km Smart 7 Normalized Delta Skews

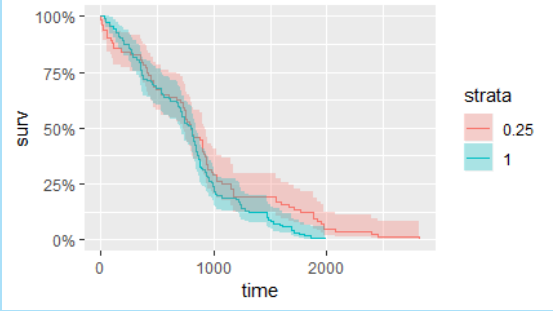

Figura 21: km Smart 192 Normalized Delta Skews

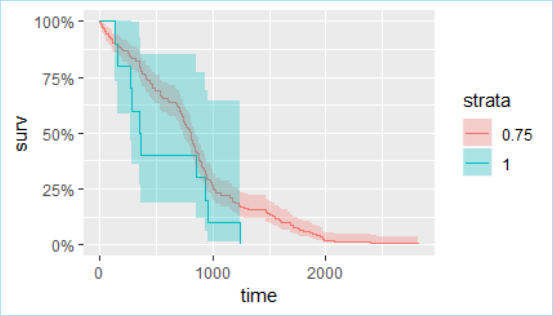

Figura 22: km Smart 193 Normalized Delta Skews

Figura 23: Gráficas

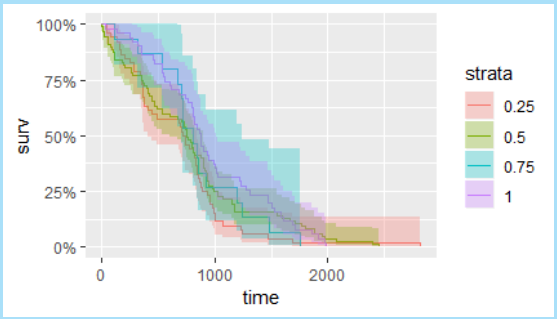

Figura 24: km Smart 194 Normalized Skews

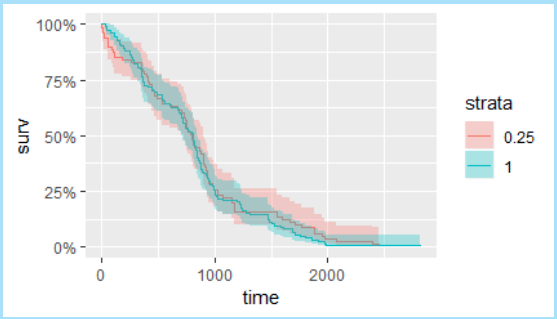

Figura 25: km Smart 197 Normalized Skews

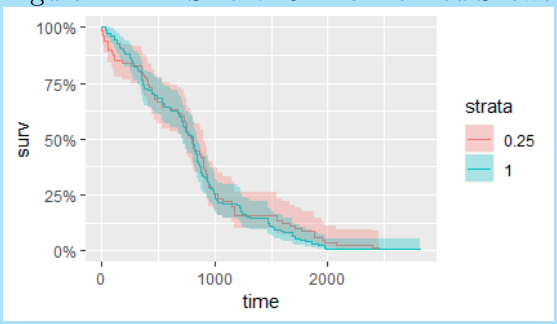

Figura 26: km Smart 198 Normalized Skews

Figura 27: Gráficas

Se aprecia de los gráficos de las Figuras 23 y27 que en solo tres de ellos poseen curvas que son notoriamente distintas. Esto fue validado por los test Lograng y el test de Peto, los cuales en conjunto señalaban que no había evidencia para señalar que las gráficas son distintas, tanto en la cola de la derecha como de la izquierda. Por lo tanto se descartarán todas estas variables, para los pasos siguientes.

Por lo tanto, para realizar la inspeccione gráfica del supuesto de riesgos proporcionales, solo nos quedamos las variables representadas en las figuras 18, 20 y 22.

c) Inspeccione gráficamente el supuesto de riesgos proporcionales.

Solución:
Para realizar una interpretación gráfica de los riesgos proporcionales se tomó en cuenta los siguientes gráficos:

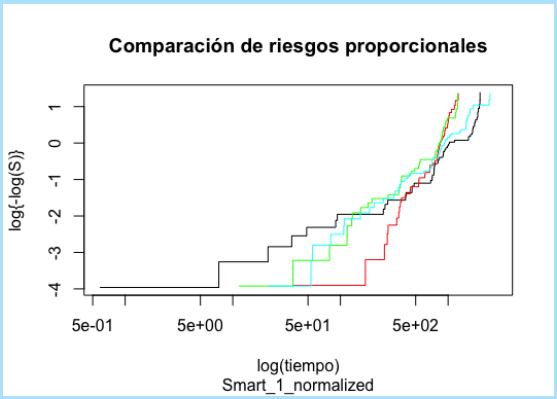

Figura 28: loglog Smart 1 Normalized

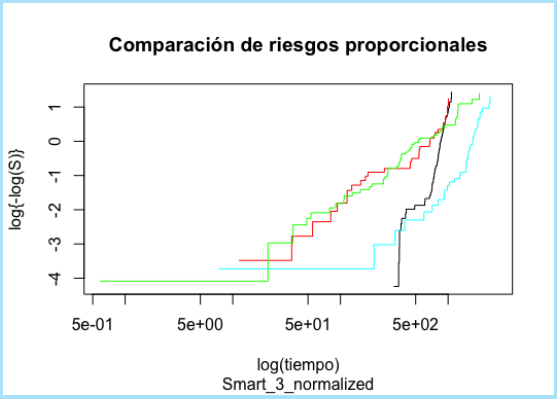

Figura 29: LogLog Smart 3 Normalized

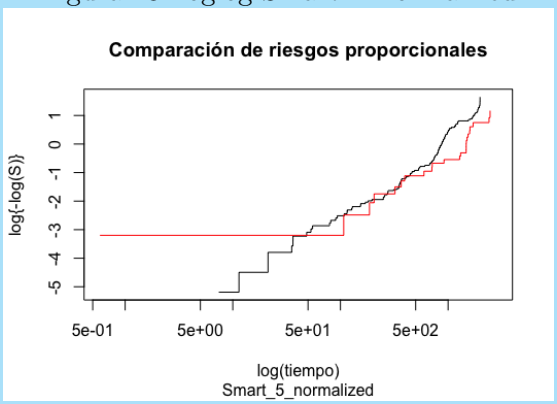

Figura 30: LogLog Smart 5 Normalized

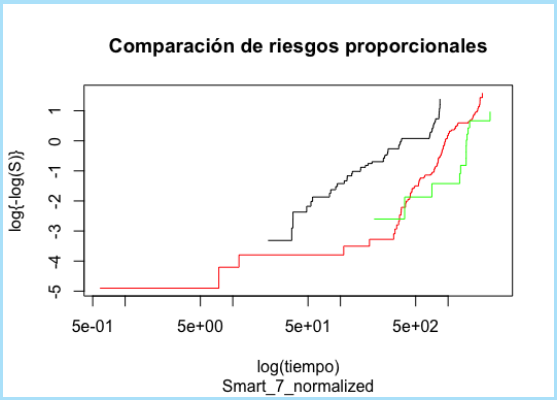

Figura 31: LogLog Smart 7 Normalized

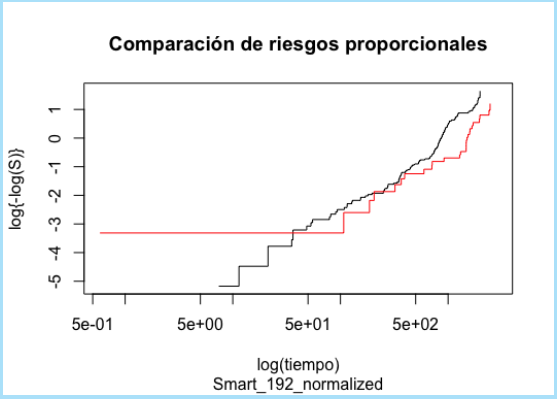

Figura 32: LogLog Smart 192 Normalized

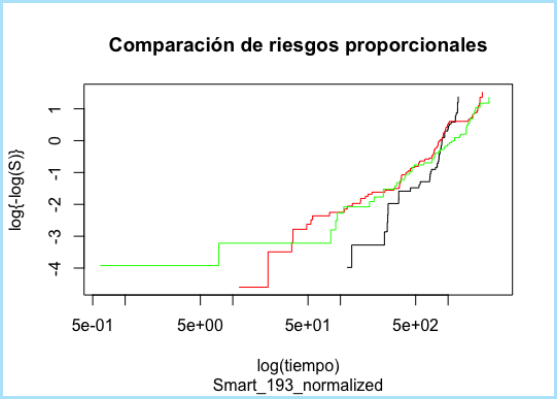

Figura 33: LogLog Smart 193 Normalized

Figura 34: Gráficas

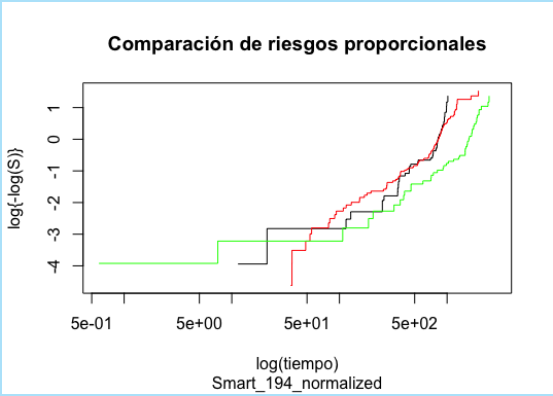

Figura 35: LogLog Smart 194 Normalized

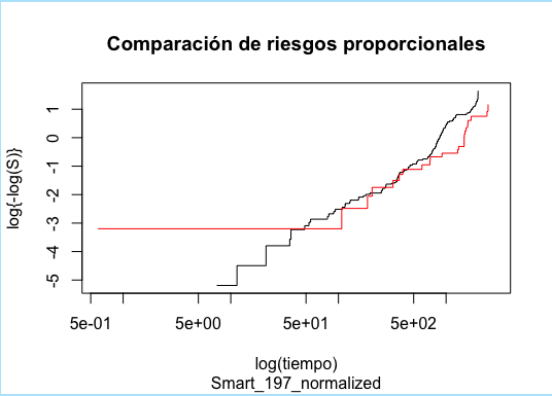

Figura 36: LogLog Smart 197 Normalized

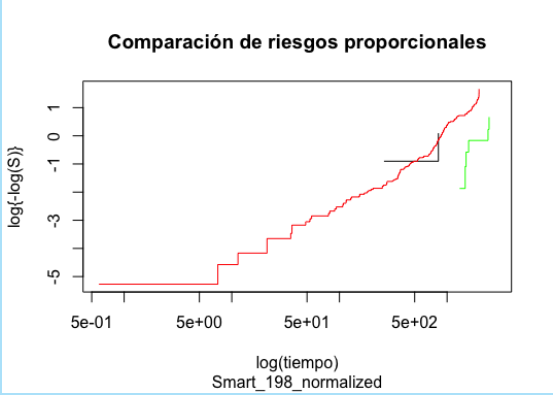

Figura 37: LogLog Smart 198 Normalized

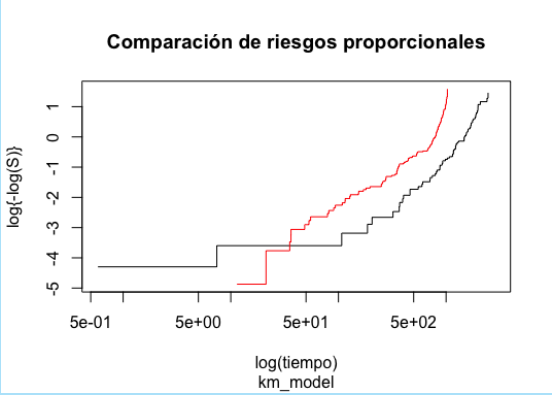

Figura 38: LogLog Model

Figura 39: Gráficas

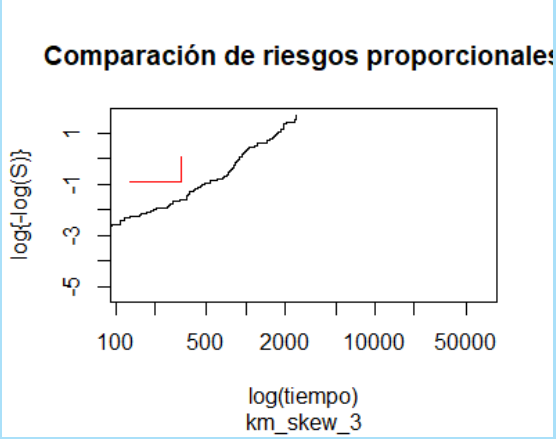

Figura 40: LogLog delta skews 3

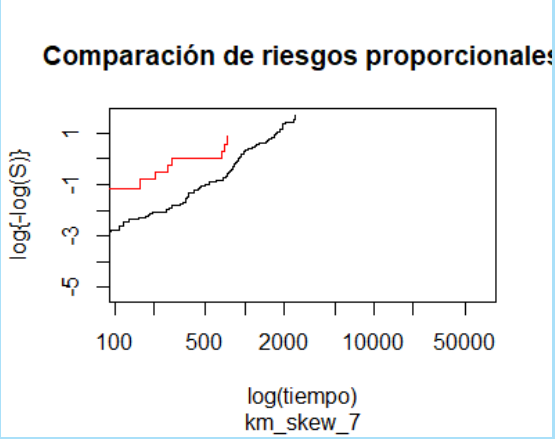

Figura 41: LogLog delta skews 7

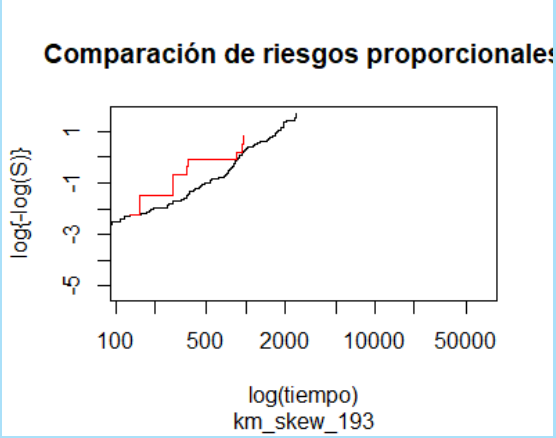

Figura 42: LogLog delta skews 193

Figura 43: Gráficas

Para la interpretación de los gráficos, se considerará que los riesgos son proporcionales en la medida en que las curvas sean paralelas. Al observar los gráficos contenidos en los cuadros 34, 39 y 43 , se puede concluir que el supuesto de proporcionalidad de los riesgos no se cumple, debido a que en ninguno de los gráficos se apreció que todas sus curvas fueran paralelas entre sí.

Por lo que, para la parte 3 de esta prueba, para poder generar los modelos se asumirá que los riesgos son proporcionales. Aún cuando se obtuvo que estos no cumplen el supuesto de proporcionalidad.

3. Parte

Modelos de Sobrevivencia.

a) Ajuste un modelo exponencial y un modelo weibull, ¿cuál de estos modelos explica mejor los tiempos de vida?, fundamente su respuesta. Interprete los resultados del modelo.

Solución:
Para comenzar, se asumirá que el supuesto de proporcionalidad de riesgo se cumple para llevar a

cabo esta parte de la prueba. Con ello en cuenta, y considerando el análisis realizado en la ref, se tomaron como variables categóricas a las variables "model", "smart_3". En cambio, las variables "smart_5","smart_7", "smart_192", "smart_194", "smart_197", "smart_198", "smart_3_skew", "smart_7_skewz "smart_193_skew"se consideraron como variables numéricas. Por otra parte, se dejaron fuera las variables, "smart_1", "smart_193", "smart_1_skew", "smart_5_skew", "smart_192_skew", "smart_194_skew", "smart_197_skew", "smart_198_skew".

El resultado de los modelos comparados por medio de las gráficas de curvas Kaplan meier es el siguiente.

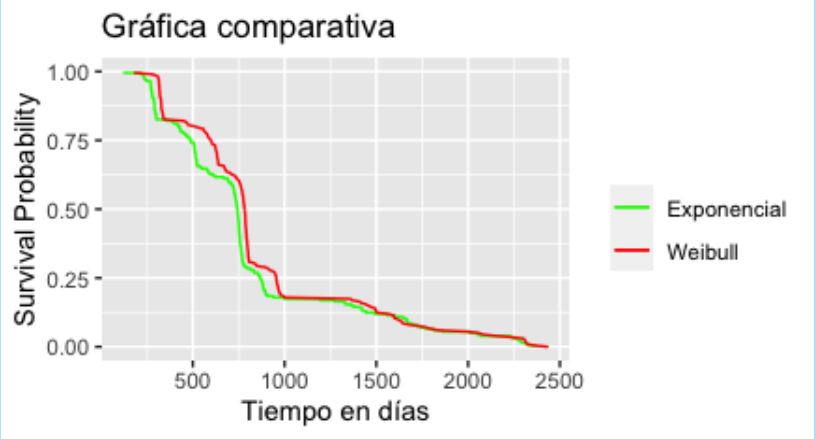

Figura 44: Exponencial versus Weibull

Modelo	Log-likelihood
Exponencial	-1545.2
Weilbull	-1480.3

Cuadro 11: Verosimilitud Exponencial y Weibull

Dado el gráfico 44 y la tabla 11, se puede concluir que el modelo que mejor se adapta a los datos es el modelo **Weibull**, dado que tiene una verosimilitud mayor y equivalente a: -1480.3, en contraste a la verosimilitud del modelo exponencial que posee un valor de -1545.2. Y en comparativa tiene una mejor curva de sobrevivencia que la Exponencial.

Para la interpretación de los resultados de los modelos además de la gráfica 44, se generó por medio de los coeficientes de cada modelo (β) los siguientes gráficos. En los cuales están presentes los valores de la tasa de riesgo asociada a cada variable, en cada uno de los modelos. Esta tasa es para las variables numéricas e^{β} y para las variables categóricas $e^{-\beta}$ debido a una particularidad de r. Por medio de las siguientes gráficas se puede observar como los modelos caracterizan a los discos como más riesgosos según sus atributos.

El primer gráfico, contiene a todas las variables de los modelos permanecieron como variables numéricas:

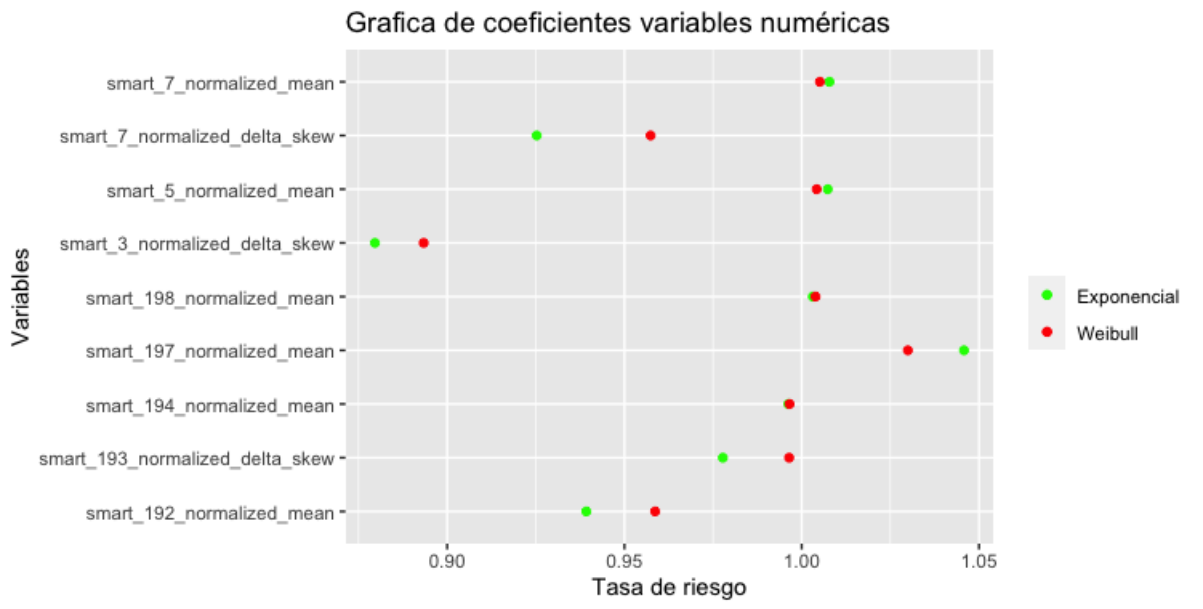

Figura 45: Tasa de Riesgo Variables numéricas

Tomando en cuenta la existencia de los puntos se aprecia que el modelo Exponencial, en general considera menos riesgosas las variables numéricas, en donde solo resalta la variable smart 197 normalized mean en donde señala que aumenta el riesgo del disco a medida que este valor es más alto, pero tomando en cuenta la gráfica 15 podemos apreciar que el riesgo debiese disminuir al aproximarse al ultimo cuartil, ya que esta curva se posiciona en un tiempo de supervivencia más alto que el 75 % de los datos inferiores. Por otro lado se aprecia que cuando se considera que al aumentar el valor unitario de una variable disminuye el riesgo (tasa de riesgo menor a 1) el modelo exponencial exagera esta disminución del riesgo entregando valores de tasas de riesgo mucho más pequeños que el modelo Weilbull.

El segundo, esta hecho en base a las variables que fueron utilizadas como categóricas:

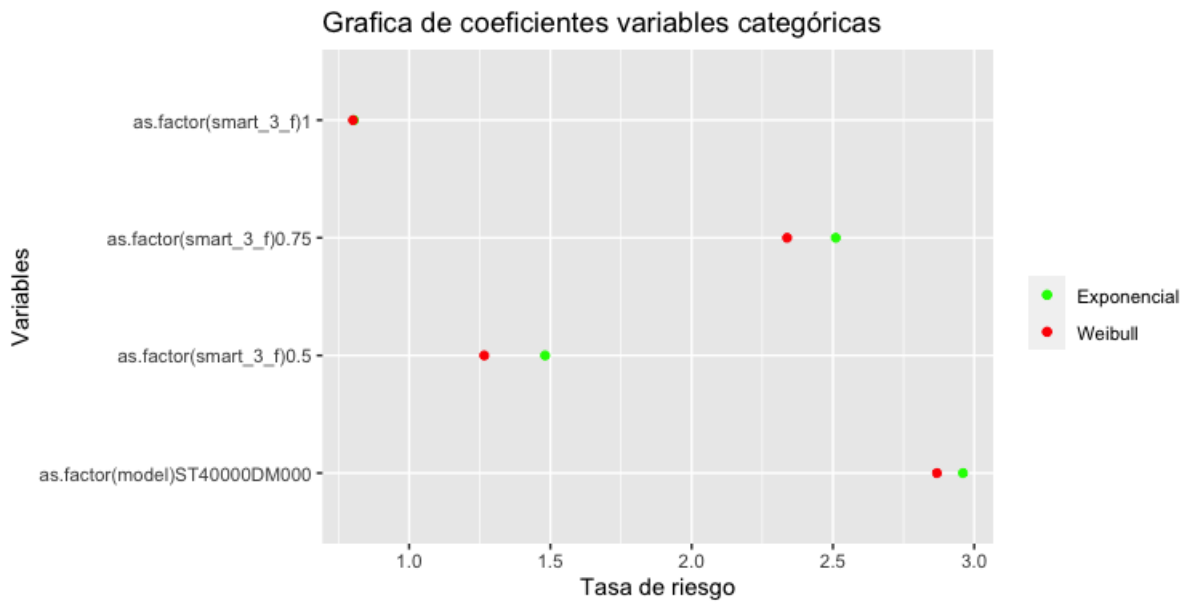

Figura 46: Tasa de Riesgo Variables categóricas

Se aprecia de la gráfica 46 que nuevamente que los modelos se comportan de forma similar a la gráfica 45, en donde el modelo Exponencial resalta los riesgos, manteniendo tasas de riesgo mucho mas altas que el modelo Weilbull. Además se aprecia que ambos modelos consideran altamente riesgosos los discos con del modelo `ST4000DM00`, siendo el modelo Exponencial el que lo considera más riesgoso y hasta 3 veces más riesgoso que cualquier otro modelo de discos(`'other'`). Tomando en cuenta la curvas de Kaplan Meier de la gráfica 7 se aprecia que considerar estos discos como más riesgosos esta en lo correcto, ya que entre todas las gráficas de KM, esta división se comportó de forma mas significativa.

En cuanto a los resultados obtenidos del modelo Weibull, el cual es considerado por la métrica de Log-likelihood mejor modelo que el Exponencial, este según la gráfica 44 es un modelo menos pesimista que el su par, manteniendo tiempos de supervivencia más altos en casi todo momento, ya que esta curva se posiciona a una altura mayor. Contrastando esta imagen con los gráficos vistos anteriormente(**??** y 46) es debido a que el modelo exponencial exagera los riesgos que presentan ciertas categorías, lo cual implica que en general consideré que los discos viven menos de lo que realmente viven.

b) Ajuste un modelo lognormal y gamma ¿en comparación al modelo que seleccionó en 6, es uno de estos modelos mejor?.

Solución:
Luego de ajustar los modelos Lognormal y Gamma, se obtuvo el siguiente resultado de los valores de Verosimilitud de los modelos, que como se explicó en el inciso anterior (**a**), sirve para determinar que modelo se ajusta mejor a los datos:

Modelo	**Log-likelihood**
LogNormal	-1559.5
Gamma	-1502.626

Cuadro 12: Verosimilitud Lognormal y Gamma

De la tabla 12, se desprende que el modelo que mejor se adapta a los datos con los que se está trabajando, fue el modelo **Gamma**, con un LogLikelihood de: $-1502{,}626$, que es mayor al obtenido por el Lognormal que fue de: $-1559{,}5$.

Por otra parte, en la siguiente tabla se presenta la comparación entre el modelo **Gamma** seleccionado y el modelo que fue elegido en la parte **a**:

Modelo	Log-likelihood
Gamma	-1502.626
Weibull	**-1480.3**

Cuadro 13: Verosimilitud Weibull y Gamma

De la tabla 13, se desprende nuevamente, que el mejor modelo para este conjunto de datos fue el modelo **Weibull**, con una verosimilitud de: $-1480{,}3$ que es mayor a $-1502{,}626$ del modelo Gamma.

c) Del modelo paramétrico seleccionado, compare las curvas de sobrevivencia en relación a la curva de Kaplan-Meier

Solución:
El siguiente gráfico compara las curvas de sobrevivencia para **Kaplan Meier** y el modelo paramétrico seleccionado en las partes **a** y **b**, que fue el modelo **Weibull**:

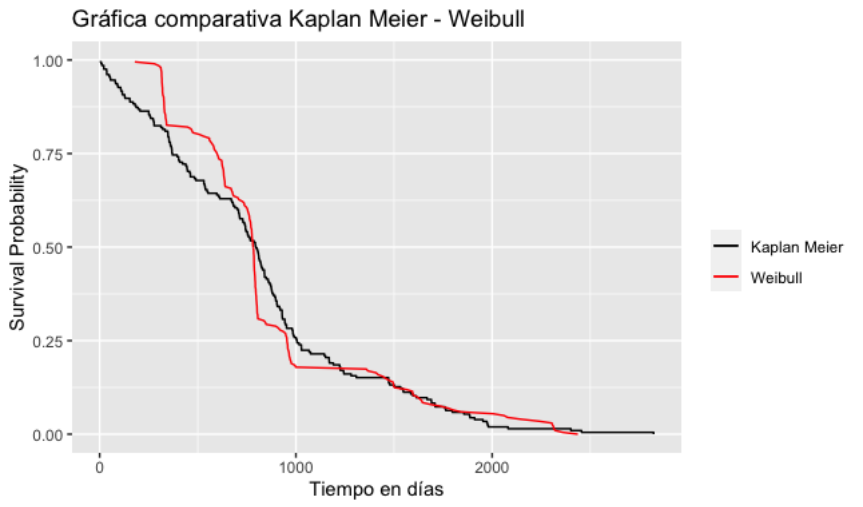

Figura 47: Kaplan Meier versus Weibull

Del gráfico 47, se desprende que el modelo Weibull logra ajustarse de una forma bastante cercana a la realidad, con algunas desviaciones menores en el inicio de la curva y otras aún menores en el medio de esta. De modo que se puede comprender, porque ha sido el mejor modelo paramétrico, al contemplar esta curva.

d) Ajuste un modelo de regresión de Cox. Compare los resultados en comparación a los modelos desarrollados previamente. Interprete los resultados en términos de los coeficientes de regresión.

Solución:

Una vez ajustado el modelo de regresión de cox, se realizó una comparación con el modelo **Weibull** seleccionado en los incisos **a** y **b**, y el modelo **Kaplan Meier**. Resultando el siguiente gráfico:

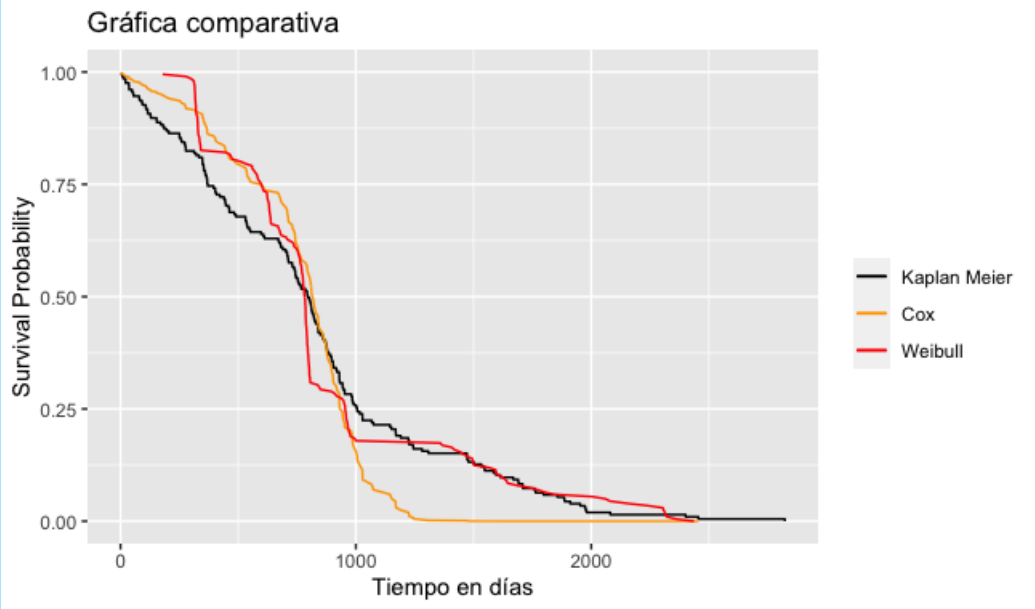

Figura 48: Cox vs Kaplan Meier - Weibull

Respecto al gráfico 48, se desprende que el modelo **Weibull**, continúa siendo el modelo que mejor se ajusta a los datos, aún cuando en el inicio de la curva el modelo de regresión de Cox parece ser levemente más acertado, en el continuo, es el modelo Weibull quien mantiene el mejor comportamiento respecto a los datos.

Modelo	**Log-likelihood**
Cox	-761.6486
Weibull	-1480.3

Cuadro 14: Verosimilitud Weibull y Cox

Tomando en cuenta el resultado de la tabla 14, se rechazaría lo observado en la gráfica 48, ya que el test de verosimilitud aplicado (**Log-likelihood**) es más certero que un análisis visual, ante esta incertidumbre se le consulto al profesor e indicó que el test aplicado tiene mayor peso en el criterio de comparación de los modelos, es por ello que se considera el modelo de cox como el mejor modelo de todos los analizados.

Como nos encontramos en una disyuntiva para determinar cual modelo se ajusta mejor a los datos, se decidió aplicar otra métrica de comparación para estos modelos, llamada Akaike's Information Criterion (AIC), que tiene como particularidad que penaliza al modelo en función de la complejidad de este:

Modelo	AIC
Cox	1549.297
Weibull	2924.416

Cuadro 15: Verosimilitud Weibull y Cox

A partir de la tabla 15, se obtiene que el menor valor de AIC fue para el modelo de regresión de Cox, con un valor de 1549,297. Con lo que se pudo determinar, cuál fue el modelo que mejor se ajusto a los datos, considerando que el método de penalización se realiza en función al número de parámetros y ambos modelos cuentan con el mismo número de estos, el mejor modelo fue el **modelo de regresión de Cox**.

Finalmente, respecto a los coeficientes de regresión obtenidos, podemos observar los siguientes gráficos respecto al riesgo relativo:

El primero, compara los coeficientes de regresión para las variables numéricas:

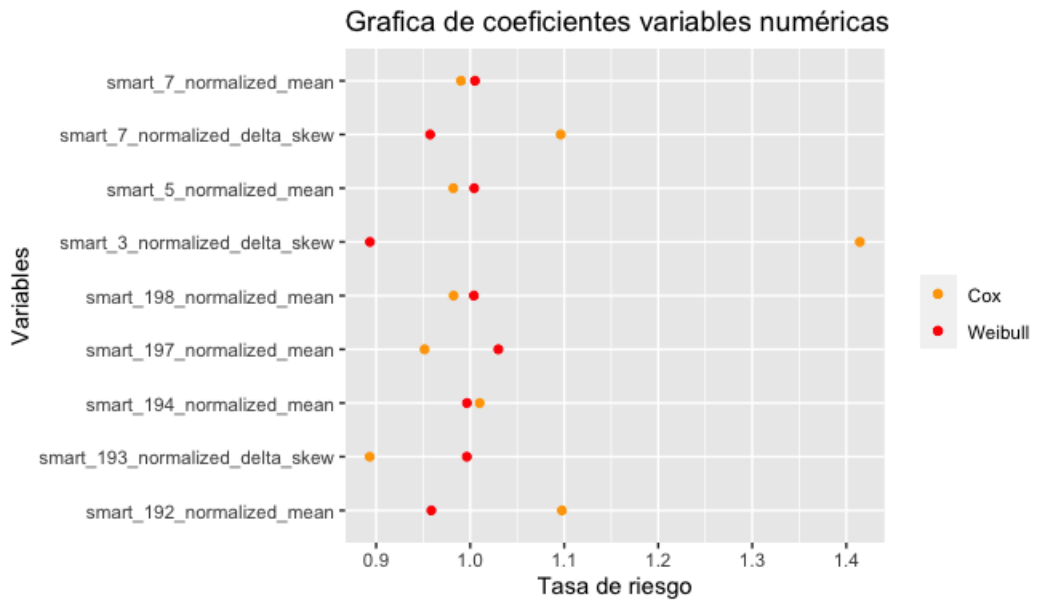

Figura 49: Tasa de riesgo Cox

Por un lado, del gráfico 49, podemos observar que no existe una diferenciación notable entre los modelos tomando en cuenta la gráfica 49 debido a que en varias variables, las tasas que asigna los modelos se cruzan constante mente, por lo cual no se aprecia un modelo que exagere los riesgos como lo ocurrido con el modelo Exponencial.

Por otro lado, se aprecia que el modelo de Cox en ciertas variables las considera altamente riesgosas, como por ejemplo considera a los discos un 40 % más riesgosos si estos cambian su distribución en la variable `smart 3 normalized` a lo largo de dos semanas en 1 unidad. Esto tomando en cuenta la tasa de riesgo asignada a la variable `smart_3_normalized_delta_skew` es de 1.4 aproximadamente. Tomando en cuenta la gráfica 18, esta afirmación esta en lo correcto, ya que el ultimo cuartil de esta

variable posee tiempos de supervivencia menores, por lo cual implica en que los discos que poseen un valor alto en esta variable tienen una tasa de riesgo más alta.

El segundo, compara los coeficientes de regresión para las variables categóricas:

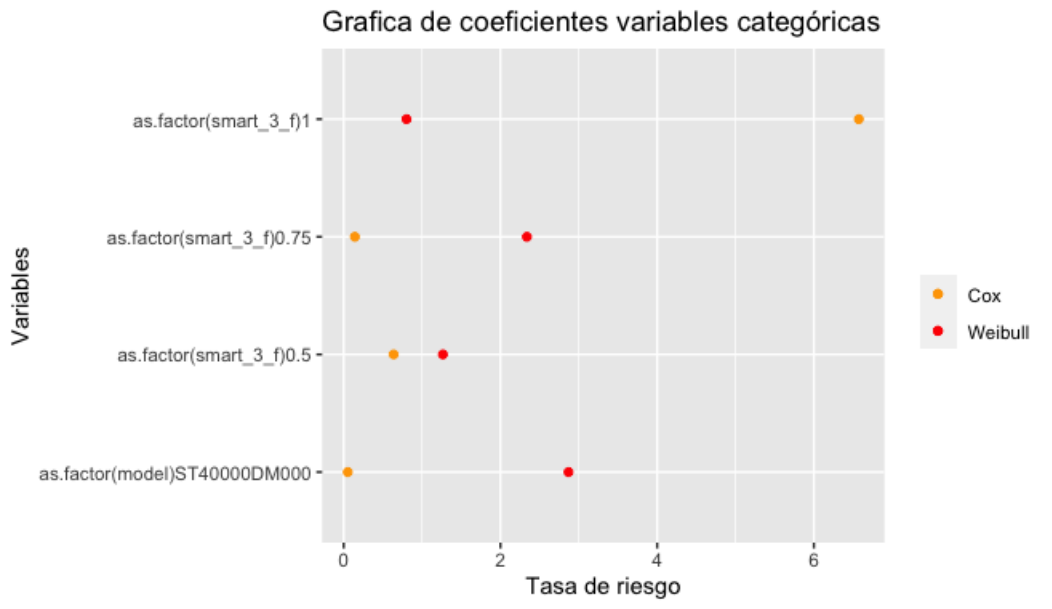

Figura 50: Tasa de riesgo Cox

A diferencia de lo ocurrido en el gráfico anterior (49), de la figura 50, se puede desprender una regla de comportamiento que diferencie a los dos modelos, en donde el modelo de Cox tiene criterios de riesgos totalmente ilógicos, ya que asigna una tasa de riesgo bajísima al modelo más riesgoso, el cual fue validado como tal con los modelos Weilbull y Exponencial, además de su comportamiento en la gráfica 7. Además, tomando en cuenta la categorización de la variable smart_3_normalized_mean, observada en la figura 9, la categoría 1 se presenta como la menos riesgosa, pero Cox la considera casi 7 veces mas riesgosa que la categoría 0.25 según la gráfica, lo que nos indica que, pareciera haber algo mal en en cálculo de las tasas de riesgo.

Por lo tanto, tomando en cuenta la gráfica 50, el cálculo de las tasas de riesgo aplicado para los modelos anteriores sobre las variables categóricas no debería ser el mismo para el modelo de Cox, esto es debido a que la función utilizada para generar el modelo es distinta, por lo tanto la interpretación de los coeficientes para las variables categóricas podría presentar diferencias. Es por esto, que se procedió a generar la siguiente gráfica calculando las tasas de riesgo para las variables categóricas de Cox de la siguiente forma e^{β} (cuando anteriormente, el cálculo se realizaba como $(e^{\beta})^{(-1)}$).

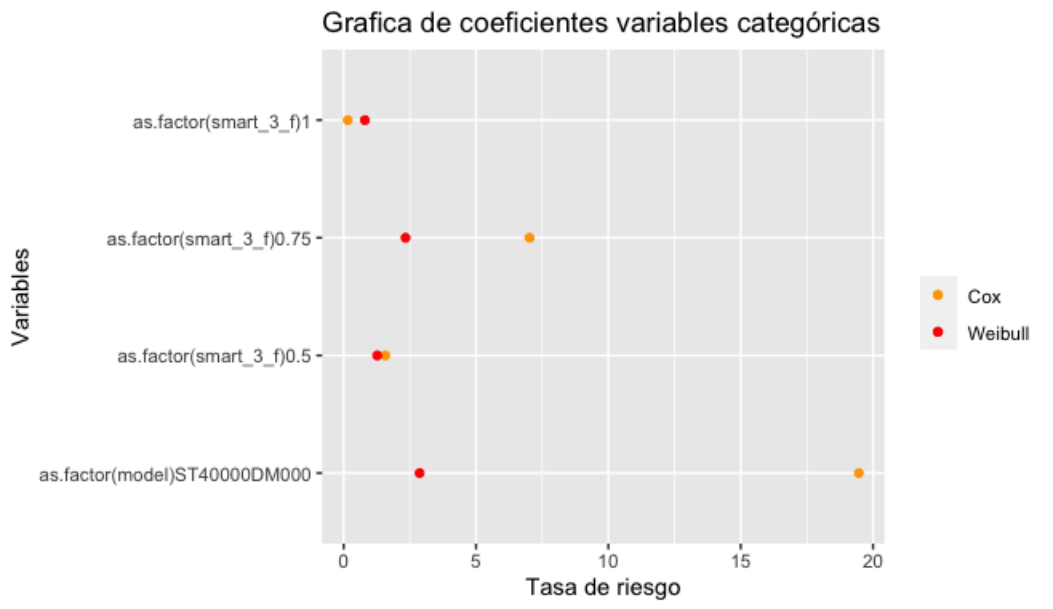

Figura 51: Tasa de riesgo Cox corregida

Se aprecia que las tasas de riesgo de Cox parecen más lógicas, además que exagera notoriamente el riesgo que representa que un disco sea del modelo `ST4000DM00` ya que lo considera hasta ocho veces más riesgoso que el modelo Weilbull, la siguiente categoría que resalta en estos ámbitos es la categoría 0.75, tomando en cuenta la gráfica 9 no debiese ser la más riesgos entre las categorías si consideramos los últimos discos que fallaron de esa categoría, lo que implica que supera con creces las categorías 0.25 y 0.5.

En conclusión, para predecir el tiempo de supervivencia de los discos, se debería utilizar el modelo de regresión de Cox, tomando en cuenta los test aplicados anteriormente. Pero si se desea verificar las tasas de riesgo sobre las variables categóricas, y tomar en cuenta las exageraciones que implica el modelo de Cox, es recomendable contrastar las tasas de riesgo con el modelo Weibull, el cual se desempeña como un modelo más ameno con respecto a las tasas de riesgo.