Trabajo Central- Big Data e Inteligencia Empresarial

Fase 1 planteamiento del proyecto.

descripción del problema:

Andika 3D realiza cotizaciones de sus productos teniendo en cuenta el peso, tamaño y cantidad de horas consumidas en la maquinaria.

Actualmente, la compañía maneja dos tamaños de su producto "MEGAPET", pero, debido al creciente interés de varios clientes, la compañía está pensando en permitirles cotizar cualquier tamaño que deseen en su producto.

En el proceso de cotización se toma muy en cuenta el peso y las horas consumidas en su proceso de manufactura, además, se realiza un prototipo real para conocer el peso final y la cantidad de horas y con base a las medidas del prototipo se establece el costo de producción. Este método de cotización toma mucho esfuerzo y tiempo, por lo que la compañía busca una forma de predecir estos costos de forma más ágil, basándose en los datos históricos de los tamaños de MegaPet que ya han realizado.

Historia de usuario (preguntas):

Basándose en la problemática, se definen las siguientes preguntas:

1. **¿Qué tamaños de MegaPet han distribuido hasta ahora?**
   - Hasta ahora hemos distribuido MegaPets de 15, 20 y 30; dentro de poco sacaremos un MegaPet de 60 cm.

2. **¿Como se pueden obtener los datos referentes al tiempo consumido en maquinaria de un producto, el peso del producto y finalmente el tamaño del producto?**
   - Manejamos un código para cada MegaPet producido (Ej. MP001), pero la información se encuentra distribuida en varias fuentes. Por una parte, tenemos un EXCEL en donde manejamos las fichas de producción, en este Excel se puede encontrar las características del producto como el tamaño en cm, por otra parte, tenemos los logs de maquinaria, en donde podemos apreciar la cantidad de filamento utilizado en mm y la cantidad de horas que tomo un archivo en imprimirse.

3. **¿Cuántos archivos de impresión componen a un MegaPet?**
   - Es variable, depende de los accesorios adicionales que haya escogido el cliente, por lo que la cantidad varía entre dos, tres, cinco y más archivos de impresión. También depende del tamaño de la pieza a imprimir, si es muy grande se divide en varias partes

4. **¿Es posible identificar que archivo de impresión le pertenece a un MP particular?**
   - Suponemos que sí, el MP esta relacionado en el nombre del archivo, desde algunos meses estamos comenzando a utilizar la nomenclatura '**MPXXX_ParteNombreMascota.gcode'** por lo que la primera parte del encabezado del archivo tendrá el código MP al que pertenece.

Fase 2 – fase investigativa y práctica.

Herramienta de analítica seleccionada:

En este caso, se decide utilizar EXCEL para la preparación de los datos y una herramienta de análisis de datos como R debido a su capacidad para visualizar datos de manera efectiva, transformarlos y su capacidad de realizar análisis exploratorios y predictivos. Además, R nos brinda facilidad en su utilización, su amplia documentación en la implementación practica de varios ejemplos similares en su comunidad la hacen una herramienta llamativa.

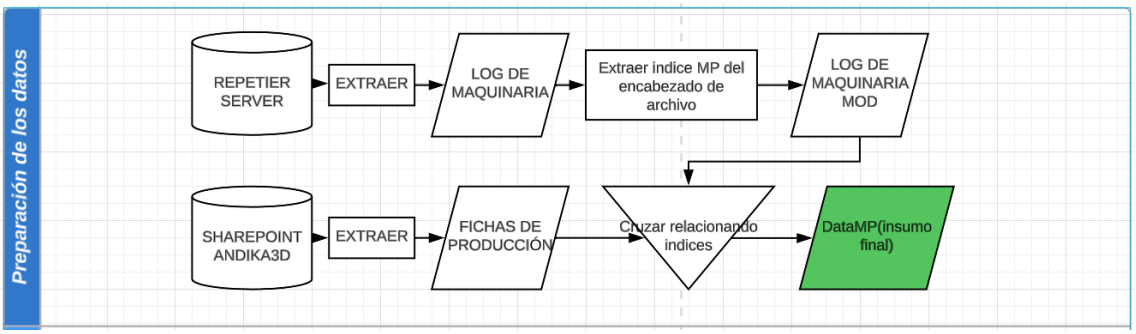
Cabe destacar que la implementación con cualquier otra herramienta como Tableu o Python nos hubiera dado buenos resultados también, pero por la curva de aprendizaje algo pronunciada en otros lenguajes, preferimos seguir con R.

Resultados obtenidos:

Preparación de los datos:

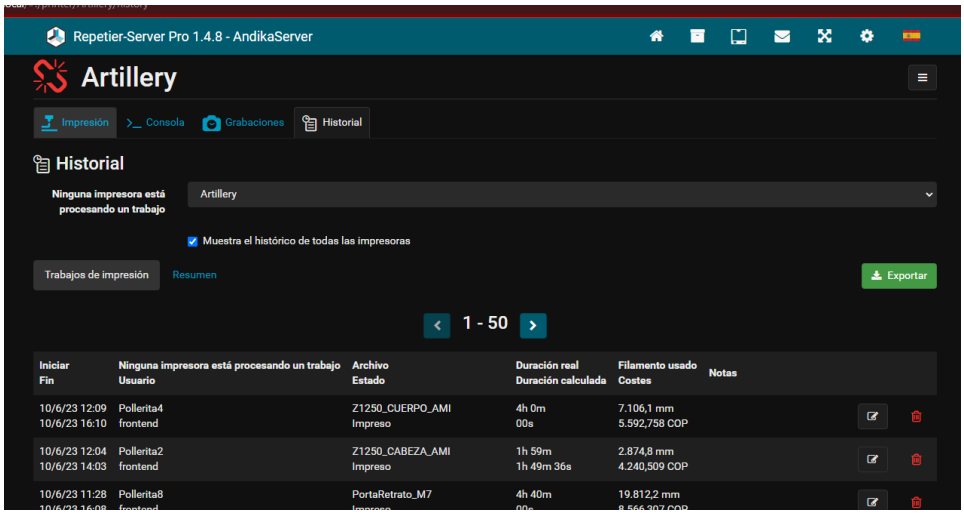
Para resolver la problemática se tiene en cuenta que la mayor carga de trabajo está en la fase de '**preparación de los datos'**, debido a que estos deben pasar primero por una serie de ajustes, teniendo en cuenta que los datos se encuentran distribuidos en diferentes fuentes no relacionables (en principio) y que el modelo de analítica que se va a aplicar utilizará los insumos de ambas.

Para esta fase se utilizo Excel, que nos permitió separar los códigos MP del encabezado del archivo, y así poder cruzar la información del log de maquinaria con las fichas de producción.

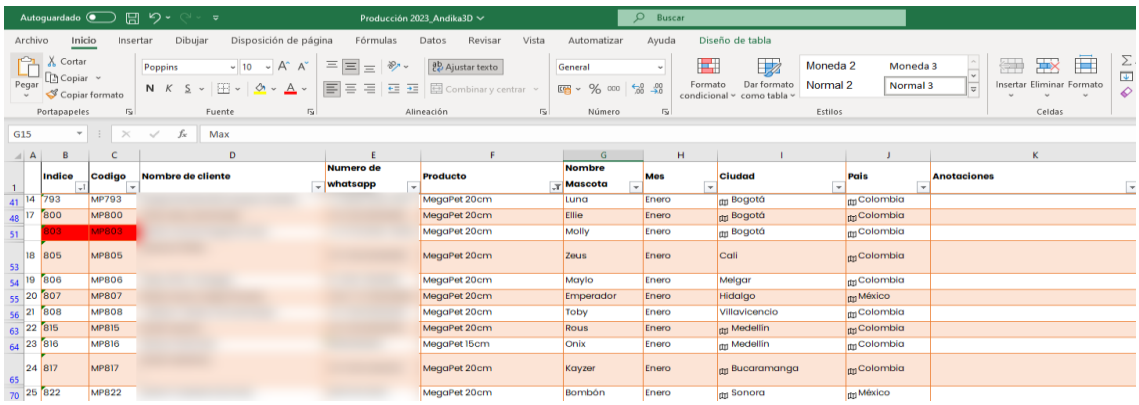


El insumo final corresponde a una tabla de 120 observaciones obtenidas a través del log de maquinaria de la compañía relacionada con las fichas de producción.
La muestra comprende un histórico desde 2022, sin embargo, para efectos prácticos se toma una muestra de las impresiones realizadas al último año 2023.

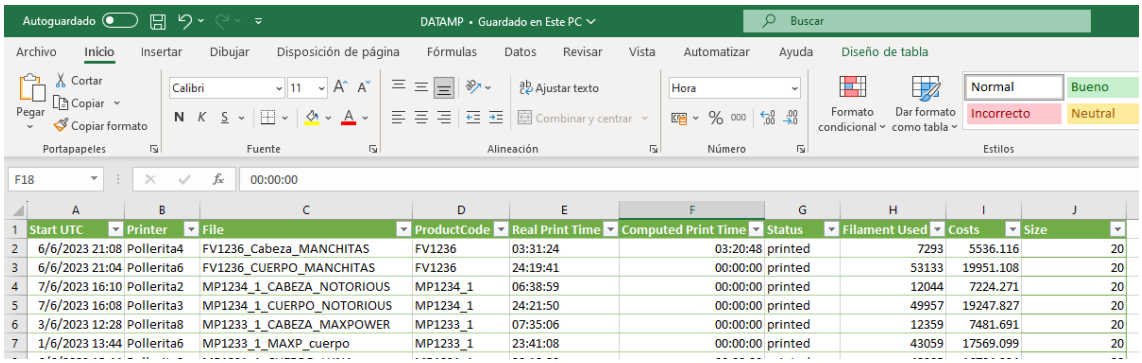
FUENTE REPETIER SERVER:



FUENTE SHAREPOINT ANDIKA3D:



INSUMO FINAL:



Después de preparar los datos, agregando las etiquetas de categorización y cruzando la información relevante entre fuentes se obtienen 120 observaciones de 10 variables:

```
#INFORMACIÓN GENERAL DEL DATASET
str(DATAMP)
```

```
tibble [120 × 10] (S3: tbl_df/tbl/data.frame)
```

las variables presentes en el dataframe corresponden a:

- **Start UTC**: Fecha y hora de inicio de la impresión en formato UTC.
- **Printer**: Nombre o identificación de la impresora.
- **File**: Nombre o identificación del archivo utilizado para la impresión.
- **ProductCode**: Código del producto.
- **Real Print Time**: Tiempo real de impresión.
- **Computed Print Time**: Tiempo calculado de impresión.
- **Status**: Estado de la impresión (printed o aborted)
- **Filament Used**: Cantidad de filamento de calibre 1.75 utilizado en milímetros.
- **Costs**: Costos asociados a la impresión (calculados por el servidor)
- **Size**: Tamaño de la impresión en centimetros.

Estas variables corresponden con los siguientes tipos de datos:

```
#Informacion del tipo de dato por columna
sapply(DATAMP, typeof)
```

```
   Start UTC            Printer               File        ProductCode     Real_Print_Time Computed Print Time
    "double"        "character"        "character"        "character"        "character"           "double"
      Status     Filament_Used              Costs               Size
 "character"           "double"           "double"           "double"
```

Notamos que la columna “**Real_Print_Time**” corresponde a una variable de tipo “carácter” en formato ‘HH:MM:SS’ por lo que para poder analizar su correlación con otras variables, y para utilizar técnicas de regresión en R tendremos que convertirla en una variable cuantificable, o en su defecto, numérica.

```
#Convertir las duraciones a segundos y luego a minutos
DATAMP$Real_Print_Time_Minutes <- period_to_seconds(hms(DATAMP$Real_
Print_Time)) / 60
```

En este caso se decide crear una columna nueva “**Real_Print_Time_Minutes**” en donde pasaremos el formato ‘HH:MM:SS’ a un formato numérico decimal de duración en minutos:

Ej.

$$27{:}51{:}04 \approx 211.40000\ minutos$$

En este punto nos interesa mantener únicamente las variables que van a ser utilizadas para el modelo de predicción y no tomar en cuenta las variables ambiguas, en este caso las variables

más representativas son el código del producto, el tiempo real de impresión en minutos, la cantidad de filamento utilizado y el tamaño del producto en cm.

Debido a que un producto puede dividirse en más de un archivo es de interés tener valores únicos en el código del producto con el fin de tener una suma totalizada del tiempo total que consumió en impresión y la cantidad total de filamento utilizado por producto, por ejemplo:

ProductCode	Real Print Time	Filament Used	Size
FV1236	210	7293	20
FV1236	1458	53133	20

ProductCode	Real Print Time	Filament Used	Size
FV1236	1668	60426	20

Para eso utilizamos la función “aggregate()” en R:

```
# Agrupar por ProductCode y calcular las sumas de Filament_Used y Re
al_Print_Time_Minutes
aggregated_data <- aggregate(cbind(Filament_Used_Sum = DATAMP$Filame
nt_Used, Real_Print_Time_Sum = DATAMP$Real_Print_Time_Minutes),
                             by = list(ProductCode = DATAMP$ProductC
ode), FUN = sum)

# Agregar la columna Size al resultado
aggregated_data$Size <- DATAMP$Size[match(aggregated_data$ProductCod
e, DATAMP$ProductCode)]
```

De esa forma obtenemos las sumatoria de los valores agrupados por el código del producto.

```
> aggregated_data
   ProductCode Filament_Used_Sum Real_Print_Time_Sum Size
1       FV1236             60426           1671.0833   20
2       MP1055             58224           1487.9333   20
3       MP1078             34533            937.4667   15
4       MP1134             80776           2207.3333   20
5       MP1136             54492           1752.5833   20
6       MP1142             61511           1876.3500   20
7       MP1146             66586           1946.6500   20
8       MP1150             52581           1479.3333   20
9       MP1152             52547           1499.1000   20
```

Modelamiento:

En base a la matriz de correlación analizada se realiza:

- Un modelo de regresión para predecir la cantidad de filamento utilizado en función del tamaño de la pieza impresa.
  En resumen, este modelo explica aproximadamente el 61.06% de la variabilidad en la cantidad de filamento utilizado, porcentaje que podría mejorarse con una creciente cantidad de datos que alimenten progresivamente el modelo analizado

- Un modelo de regresión para predecir la cantidad de tiempo real utilizado en función del tamaño de la pieza impresa.

En resumen, este modelo explica aproximadamente el 21.36% de la variabilidad en el tiempo real de impresión. Es importante tener en cuenta que este valor es relativamente bajo, lo que indica que el tamaño por sí solo no explica la mayor parte de la variabilidad en el tiempo real de impresión.

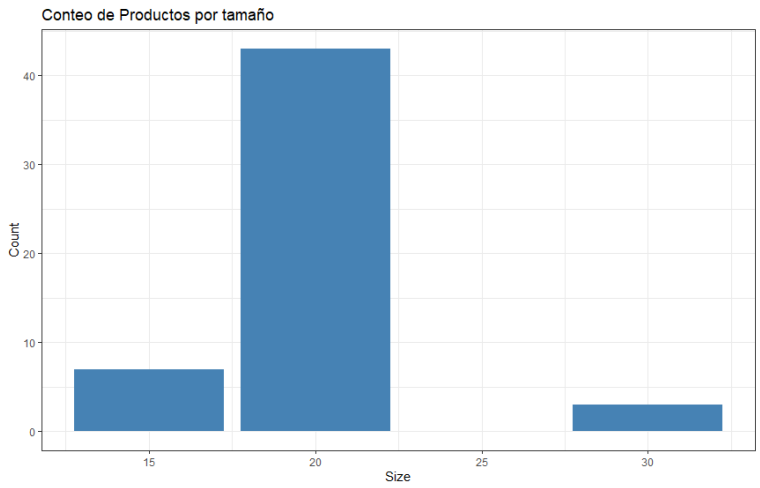
Teniendo estos resultados en cuenta, la compañía utilizara el modelo de cantidad de filamento consumido como un factor significativo para calcular los costos, en menor medida, utilizara el modelo de tiempo real consumido.

Fase 3 – Representación de la información.

Con nuestros datos preparados, realizamos un conteo de los productos por tamaño:

```
# Gráfico de barras - conteo por tamaño (Size)
ggplot(aggregated_data, aes(x = Size)) +
  geom_bar(fill = "steelblue") +
  labs(x = "Size", y = "Count") +
  ggtitle("Conteo de Productos por tamaño") +
  theme_bw()
```

Análisis gráfico



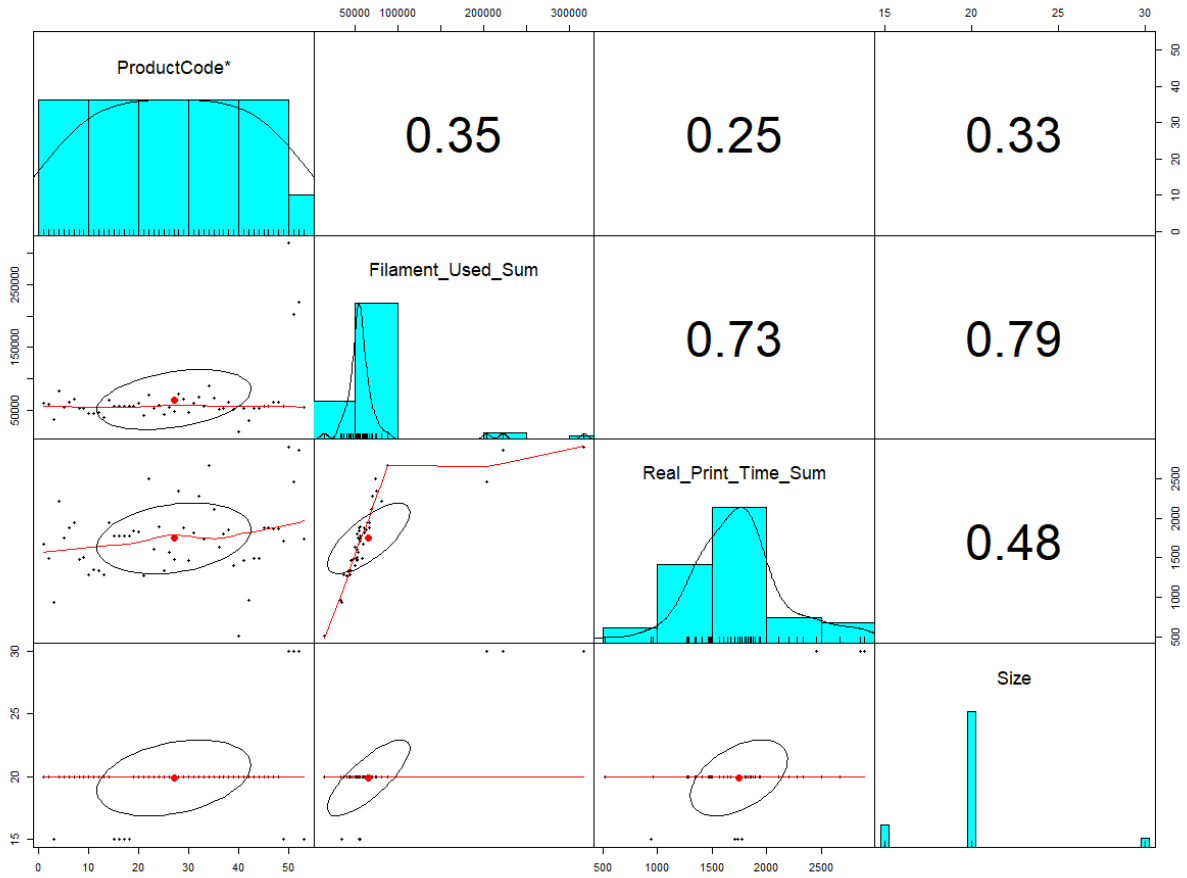
La mayoría de los productos tienen un tamaño de 20 cm. Esta es la categoría más común y cuenta con un alto número de productos realizados, los productos de 15 y 30 son considerablemente mas bajos. La compañía plantea ir suministrando mas escenarios al modelo de predicción con el fin de aumentar su precisión.

Generamos un diagrama en donde, en conjunto, visualizamos la dispersión de todas las combinaciones posibles de las variables en el conjunto de datos. Además, el grafico también

proporciona los valores de correlación para cada par de variables y nos muestra la distribución de cada variable en la diagonal del gráfico.

```
# Diagrama de dispersión, valores de correlación y distribuciones de cada variable

pairs.panels(aggregated_data,gap=0)
```



Basándonos en los anteriores conjuntos de gráficos de podemos hacer las siguientes conclusiones preliminares:

- No se observa una relación clara entre "Filament_Used_Sum" y "Size". Los puntos están dispersos y no siguen una tendencia específica.

- Hay una correlación positiva fuerte entre la variable "Filament_Used_Sum" y la variable "Size" con un valor de correlación de 0.786. Esto sugiere que a medida que aumenta el tamaño de impresión, la cantidad de filamento utilizado tiende a ser mayor.

- No se observa una relación clara y directa entre el tamaño "Size" y el tiempo de impresión “Real_Print_Time_Sum” en el gráfico de dispersión. Los puntos parecen dispersos y no siguen una tendencia específica.

- También hay una correlación positiva moderada entre la variable "Filament_Used_Sum" y la variable "Real_Print_Time_Sum" con un valor de correlación de 0.731. Esto indica que a medida que aumenta el tiempo real de impresión, la cantidad de filamento utilizado tiende a aumentar

- Por otro lado, la variable "Real_Print_Time_Sum" y la variable "Size" muestran una correlación positiva más débil, con un valor de correlación de 0.478. Esto sugiere que el tiempo real de impresión no está tan directamente relacionado con el tamaño de impresión como lo está con la cantidad de filamento utilizado.

- La variable “Real_Print_Time_Sum” parece seguir una distribución normal

Modelo de regresión lineal simple.

Para ver el performance en la práctica, construimos un modelo de regresión lineal para “Filament_Used”, “Size” y para “Real_Print_Time”, “Size”

```
# Construir un modelo de regresión para Filament_Used y Size
filament_model <- lm(Filament_Used_Sum ~ Size, data = aggregated_dat
a)
summary(filament_model)
```

```
Call:
lm(formula = Filament_Used_Sum ~ Size, data = aggregated_data)

Residuals:
   Min     1Q Median     3Q    Max
-52665 -15539  -9246   6717 124488

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -181476      27580  -6.580 2.51e-08 ***
Size           12447       1370   9.085 3.06e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 29850 on 51 degrees of freedom
Multiple R-squared:  0.6181,   Adjusted R-squared:  0.6106
F-statistic: 82.53 on 1 and 51 DF,  p-value: 3.064e-12
```

Del resumen del modelo de regresión para predecir la cantidad de filamento utilizado en función del tamaño, podemos sacar las siguientes conclusiones:

1. Coeficientes: El intercepto tiene un valor estimado de -181,476 y el coeficiente de tamaño tiene un valor estimado de 12,447. Estos coeficientes indican que, en promedio, por cada unidad de aumento en el tamaño, se espera un aumento de 12,447 en la cantidad de filamento utilizado. Además, ambos coeficientes son estadísticamente significativos ($p < 0.001$), lo que sugiere una relación significativa entre el tamaño y la cantidad de filamento utilizado.

2. R-cuadrado ajustado: El valor del R-cuadrado ajustado es 0.6106, lo que indica que el modelo explica aproximadamente el 61.06% de la variabilidad en la cantidad de filamento utilizado. Esto sugiere que el tamaño del objeto impreso es un predictor relevante para explicar las variaciones en la cantidad de filamento utilizado.

3. Estadísticas de ajuste: El valor del estadístico F es 82.53, con un p-valor extremadamente bajo (3.064e-12), lo que indica que el modelo de regresión es globalmente significativo y proporciona un ajuste significativamente mejor que un modelo sin variables predictoras. Además, el error estándar residual es de 29,850, lo que indica la dispersión promedio de las observaciones con respecto a la línea de regresión.

En resumen, el modelo de regresión muestra que el tamaño del objeto impreso tiene una relación significativa con la cantidad de filamento utilizado. A medida que aumenta el tamaño, se espera un aumento en la cantidad de filamento utilizado, de acuerdo con los coeficientes del modelo.

```
# Construir un modelo de regresión para Real_Print_Time y Size
time_model <- lm(Real_Print_Time_Sum ~ Size, data = aggregated_data)
summary(time_model)
```

```
Call:
lm(formula = Real_Print_Time_Sum ~ Size, data = aggregated_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1236.75  -272.25    58.97   309.33   917.89

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   337.00     366.12   0.920 0.361668
Size           70.73      18.19   3.889 0.000293 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 396.2 on 51 degrees of freedom
Multiple R-squared:  0.2287,   Adjusted R-squared:  0.2136
F-statistic: 15.12 on 1 and 51 DF,  p-value: 0.0002926
```

De la salida del resumen del modelo de regresión lineal para la variable "Real_Print_Time_Sum" en función del tamaño, podemos obtener las siguientes conclusiones:

1. Coeficientes: El coeficiente para la variable "Size" es 70.73. Esto significa que, en promedio, por cada unidad de aumento en el tamaño, se espera un aumento de aproximadamente 70.73 en el tiempo real de impresión. El intercepto (337.00) indica el valor esperado del tiempo real de impresión cuando el tamaño es igual a cero, pero no es relevante en este caso ya que no hay tamaños inferiores a cero.

2. Significancia estadística: El coeficiente para la variable "Size" es estadísticamente significativo, con un valor p muy bajo (0.000293). Esto sugiere que el tamaño tiene un efecto significativo en el tiempo real de impresión.

3. R-cuadrado ajustado: El valor del R-cuadrado ajustado es 0.2136, lo que significa que el modelo explica aproximadamente el 21.36% de la variabilidad en el tiempo real de impresión. Es importante tener en cuenta que este valor es relativamente bajo, lo que indica que el tamaño por sí solo no explica la mayor parte de la variabilidad en el tiempo real de impresión.
4. Residuos: Los residuos tienen una media cercana a cero y no muestran patrones evidentes en función del tamaño, lo cual es una buena señal de que el modelo captura la relación lineal entre las variables.

En resumen, el tamaño tiene un efecto significativo en el tiempo real de impresión, pero solo explica una pequeña parte de su variabilidad. Es posible que otros factores no incluidos en el modelo también influyan en el tiempo real de impresión.