

UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

Maestria en Ciencia de Datos.

Metodos Estadisticos Multivariados Reporte Estadistico

MET.Rosa Isela Hernández Zamora

Alumnos: Jesus Emmanuel Ramos Davila
Marco Antonio Obregon Flores
Matricula: 1439401, 1723556

Fecha entrega: 03/28/2023

Índice general

Introducción	2
Análisis descriptivo del conjunto de datos	3
Análisis exploratorio	3
Análisis de Factores	16
Análisis de Componentes Principales	21
Conclusiones	26
Análisis exploratorio.	26
Análisis de Factores	26
Análisis de Componentes Principales	26
Referencias	27

Introducción

El presente informe tiene como objetivo utilizar técnicas estadísticas multivariadas para analizar un conjunto de datos que contiene múltiples mediciones realizadas a un motor síncrono de imán permanente (PMSM). En específico, se utilizarán el análisis de componentes principales (PCA) y el análisis de factores para reducir la dimensionalidad del conjunto de datos y descubrir patrones y relaciones entre las variables. Estas técnicas se aplicarán en R Studio, utilizando los conocimientos adquiridos en el curso de Métodos Estadísticos Multivariados. El conjunto de datos contiene 13 variables distintas, incluyendo la temperatura del refrigerante del motor, la velocidad del motor, el par del motor y la corriente en el eje d del motor, entre otros. El análisis de componentes principales permitirá simplificar el conjunto de datos encontrando mezclas de variables que describen la mayor parte de la variación en los datos, mientras que el análisis de factores buscará reducir la dimensionalidad de los datos para explicar el máximo de información contenida en ellos.

El análisis del motor síncrono de imán permanente es importante para entender su funcionamiento y optimizar su rendimiento. El objetivo del análisis multivariado es identificar patrones y relaciones entre las variables, lo que puede llevar a descubrir factores importantes que afectan el rendimiento del motor. Al reducir la dimensionalidad de los datos mediante el uso de técnicas como el análisis de componentes principales y el análisis de factores, se puede obtener una mejor comprensión de la estructura subyacente de los datos y reducir la complejidad de la información. Esto puede permitir una mejor visualización de los patrones y relaciones, lo que puede conducir a una mejor identificación de los factores clave que afectan el rendimiento del motor. Además, la reducción de la dimensionalidad también puede ayudar a simplificar el análisis y mejorar la eficiencia del procesamiento de datos.

Los datos utilizados en este análisis son reales y fueron obtenidos de Kirgiz, W. (2021). Electric Motor Temperature. Los datos se pueden encontrar en <https://www.kaggle.com/datasets/wkirgsn/electric-motor-temperature> y contienen mediciones realizadas a un motor síncrono de imán permanente (PMSM). Es importante destacar que no se eliminaron valores atípicos o faltantes en los datos.

Además, se estandarizaron los datos antes de realizar el análisis, lo que significa que se convirtieron todas las variables a la misma escala para que tengan una media de cero y una desviación estándar de uno. Esto se hizo para que las variables se puedan comparar directamente entre sí y para evitar que una variable tenga más peso en el análisis solo porque tiene valores más grandes.

En resumen, este informe utilizará técnicas estadísticas avanzadas para analizar un conjunto de datos complejo y encontrar patrones y relaciones entre las variables del motor síncrono de imán permanente.

Análisis descriptivo del conjunto de datos

Los registros corresponden a mediciones realizadas a un motor síncrono de imán permanente (PMSM), los cuales fueron muestreados a una frecuencia de 2 Hz. El conjunto de datos contiene múltiples sesiones de medición, las cuales se pueden distinguir por el identificador de perfil (`profile_id`) y tienen una duración variable de entre una y seis horas. En total, se registraron 185 horas de operación del motor.

El dataset utilizado en este análisis contiene un total de 1,330,816 mediciones realizadas al motor síncrono de imán permanente. Este es un conjunto de datos bastante grande que requiere técnicas estadísticas multivariadas avanzadas para su análisis y comprensión. La cantidad de mediciones en este conjunto de datos proporciona una gran cantidad de información sobre el comportamiento del motor, lo que puede ser útil para identificar patrones y relaciones complejas entre las variables y optimizar su rendimiento.

Tabla 1: Variables del conjunto de datos

Variable	Descripción
<code>u_q</code>	Tensión en el eje q del motor
<code>coolant</code>	Temperatura del refrigerante del motor
<code>stator_winding</code>	Temperatura del devanado del estator del motor
<code>u_d</code>	Tensión en el eje d del motor
<code>stator_tooth</code>	Temperatura del diente del estator del motor
<code>motor_speed</code>	Velocidad del motor
<code>i_d</code>	Corriente en el eje d del motor
<code>i_q</code>	Corriente en el eje q del motor
<code>pm</code>	Temperatura del imán permanente del motor
<code>stator_yoke</code>	Temperatura del yugo del estator del motor
<code>ambient</code>	Temperatura ambiente durante la medición
<code>torque</code>	Par del motor
<code>profile_id</code>	Identificador de la sesión de medición

Cabe destacar que el motor es excitado por ciclos de conducción diseñados a mano, que establecen una velocidad de referencia y un par de referencia. Las corrientes y voltajes en coordenadas d/q son resultado de una estrategia de control estándar que intenta seguir la velocidad y el par de referencia, y las variables de velocidad y torque son las cantidades resultantes logradas por esa estrategia, derivadas de las corrientes y voltajes establecidos. La mayoría de los ciclos de conducción corresponden a caminatas aleatorias en el plano velocidad-par, con el fin de imitar ciclos de conducción del mundo real de manera más precisa que las excitaciones y rampas de subida y bajada constantes.

Análisis exploratorio

En esta sección se analizarán los histogramas y boxplot de las variables, para esta sección se incluirán pruebas de normal univariada y multivariada así como grafica de correlación de pearson a fin de encontrar cuales variables se relacionan más con otras. Para esta y demás secciones se omitirá una de las variables mostradas de la sección anterior la cual es **`profile_id`** la cual no es una medición de nuestros datos y solo identifica la observación.

Histogramas

Histograma `u_q` (Tensión en el eje q del motor)

Histograma de Tensión en el eje q del motor

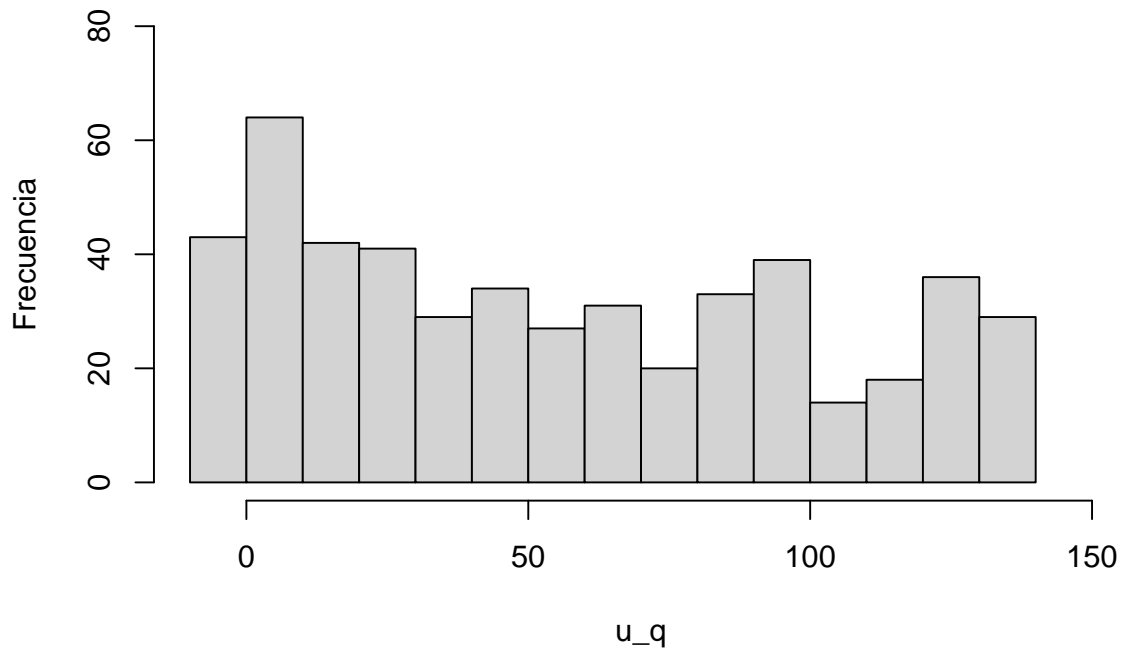


Fig 1: Histograma u_q

Observaciones: podemos observar que no se muestra una curva como una distribución de tipo normal, asemeja más a una distribución de tipo uniforme.

Histograma coolant (Temperatura del refrigerante del motor)

Histograma de temperatura del refrigerante del motor

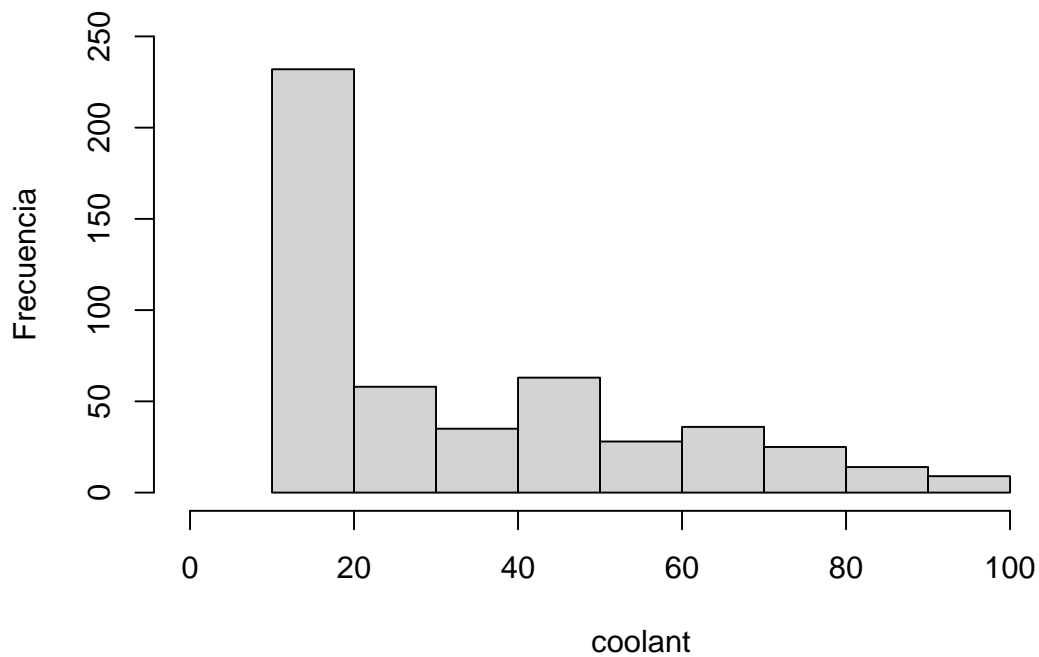


Fig 2: Histograma coolant

Observaciones: *podemos observar que no se muestra una curva como una distribución de tipo normal, asemeja más a una distribución de tipo exponencial.*

Histograma stator_winding (Temperatura del devanado del estator del motor)

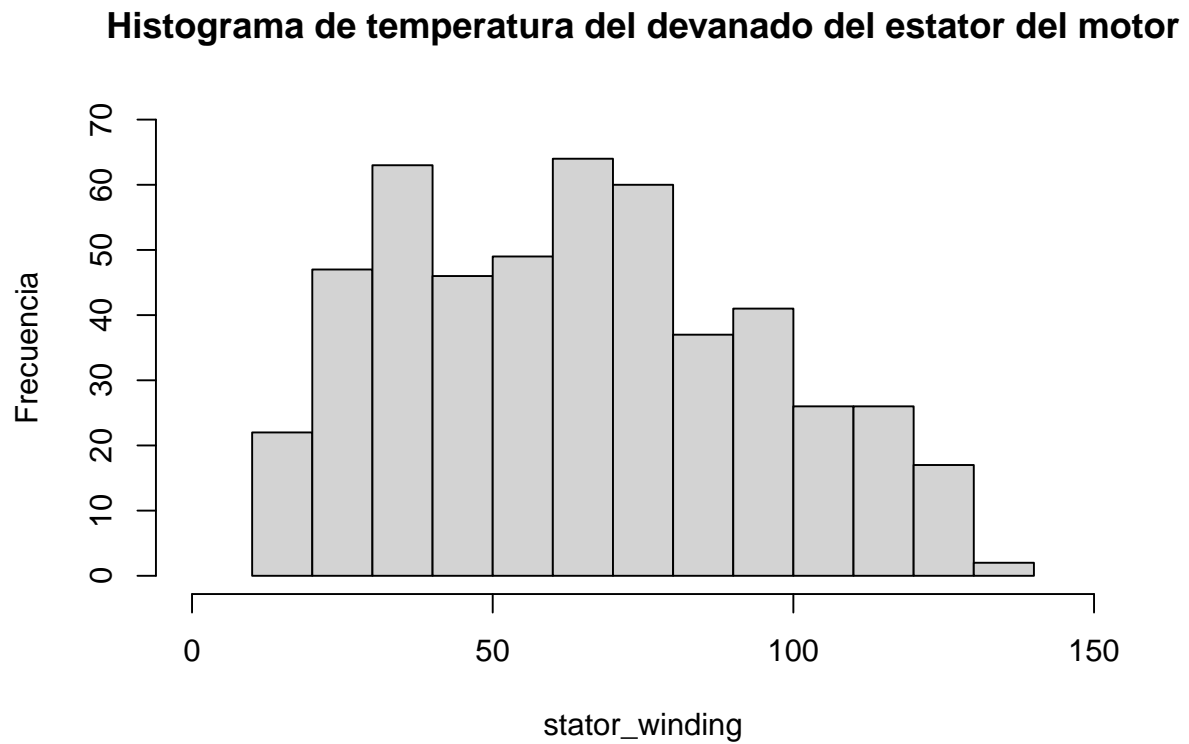


Fig 3: Histograma stator_winding

Observaciones: *Se observa en la grafica que esta variable podria seguir una distribucion de tipo normal, de cualquier manera se realizaran pruebas de normalidad univariada a fin de observar cual variable sigue una distribucion de tipo normal.*

Histograma u_d (Tensión en el eje d del motor)

Histograma de Tensión en el eje d del motor

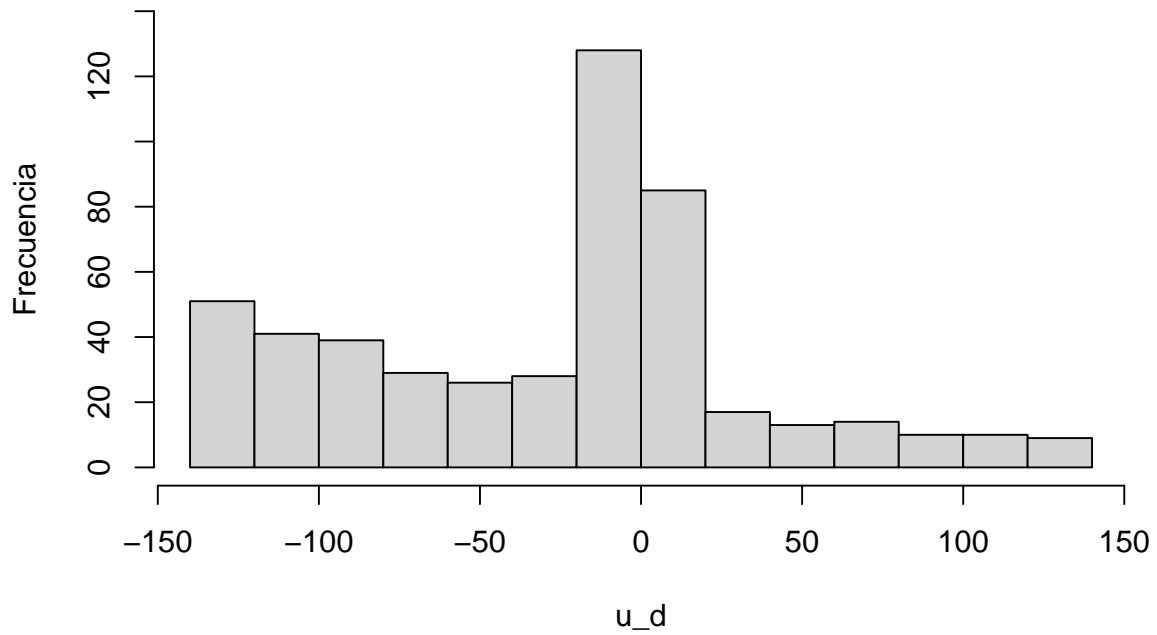


Fig 4: Histograma u_d

Observaciones: Podemos observar que no se muestra una curva como una distribución de tipo normal, se observa solo 2 barras con una gran cantidad de observaciones, se revisara en la sección de **vector de promedios** donde se ubican la media de los datos para esta variable.

Histograma stator_tooth (Temperatura del diente del estator del motor)

Histograma de temperatura del diente del estator del motor

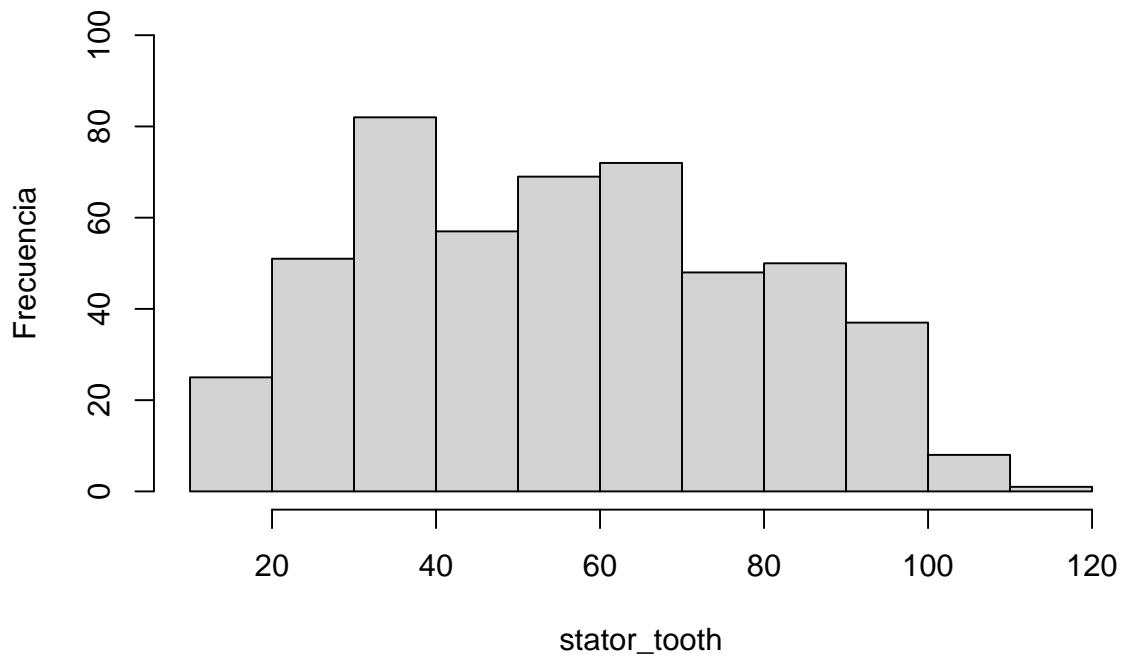


Fig 5: Histograma stator_tooth

Observaciones: Podemos observar que esta variable podría seguir una distribución de tipo normal dado los extremos y una forma un poco ligera de campana similar a la normal.

Histograma motor_speed (Velocidad del motor)

Histograma de Velocidad del motor

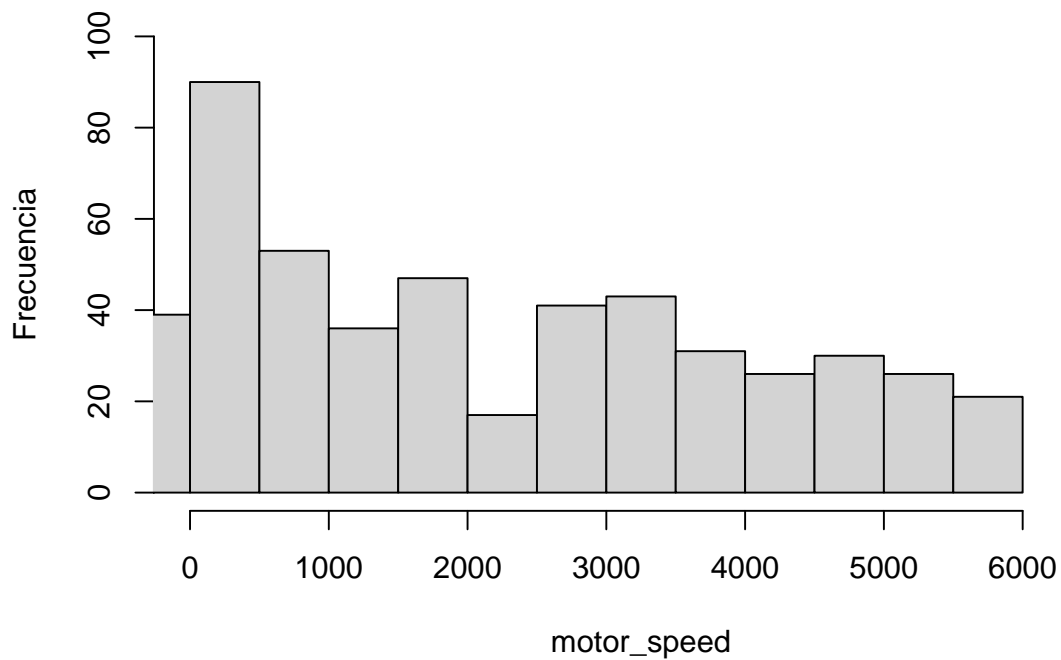


Fig 6: Histograma motor_speed

Observaciones: Podemos observar que no se muestra una curva como una distribución de tipo normal.

Histograma i_d (Corriente en el eje d del motor)

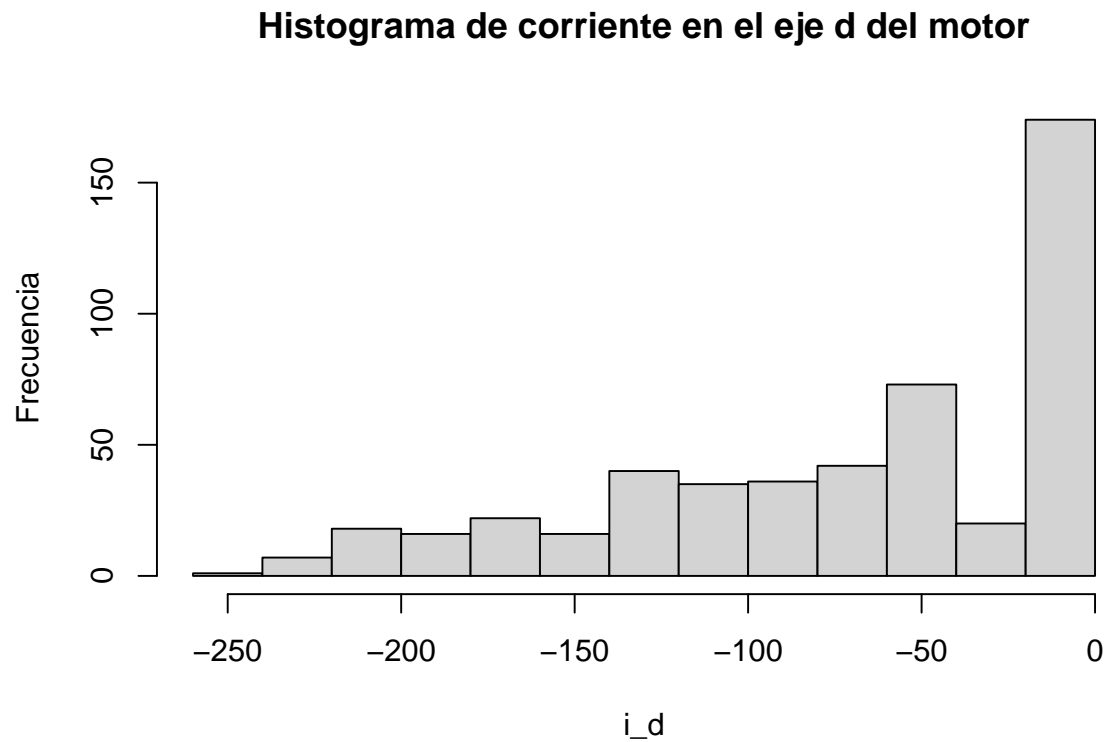


Fig 7: Histograma i_d

Observaciones: Podemos observar que esta variable no sigue una distribución normal, se asemeja más a una distribución de tipo exponencial.

Histograma i_q (Corriente en el eje q del motor)

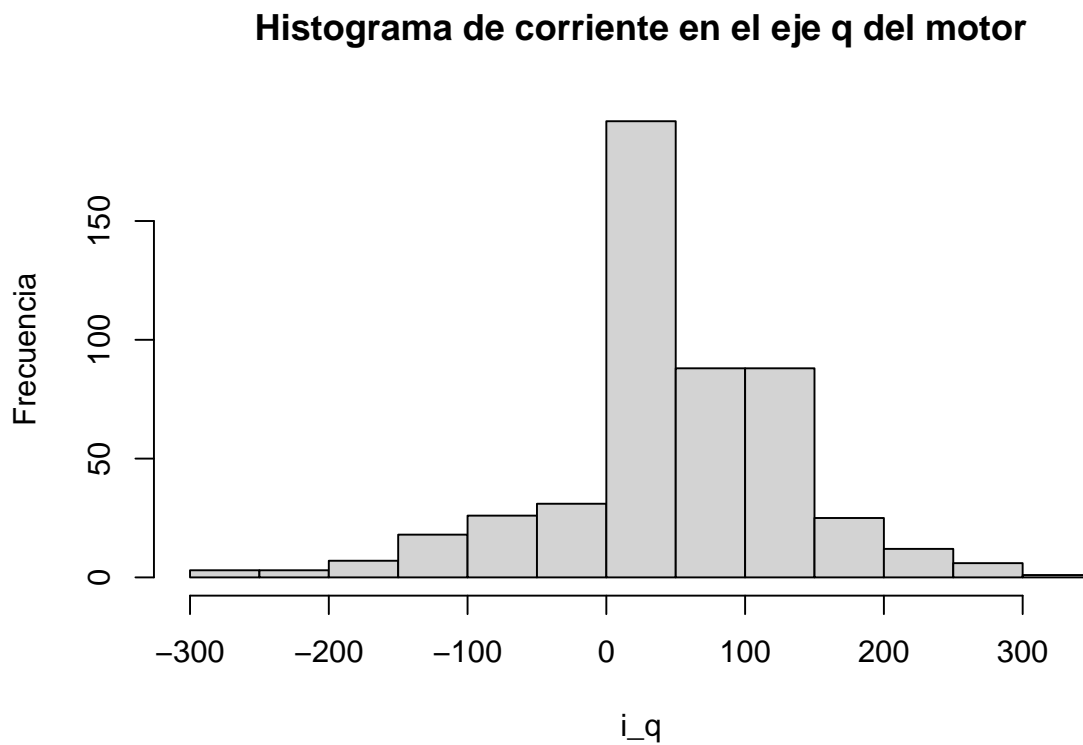


Fig 8: Histograma i_q

Observaciones: Podemos observar que esta variable no sigue una distribución normal.

Histograma pm (Temperatura del imán permanente del motor)

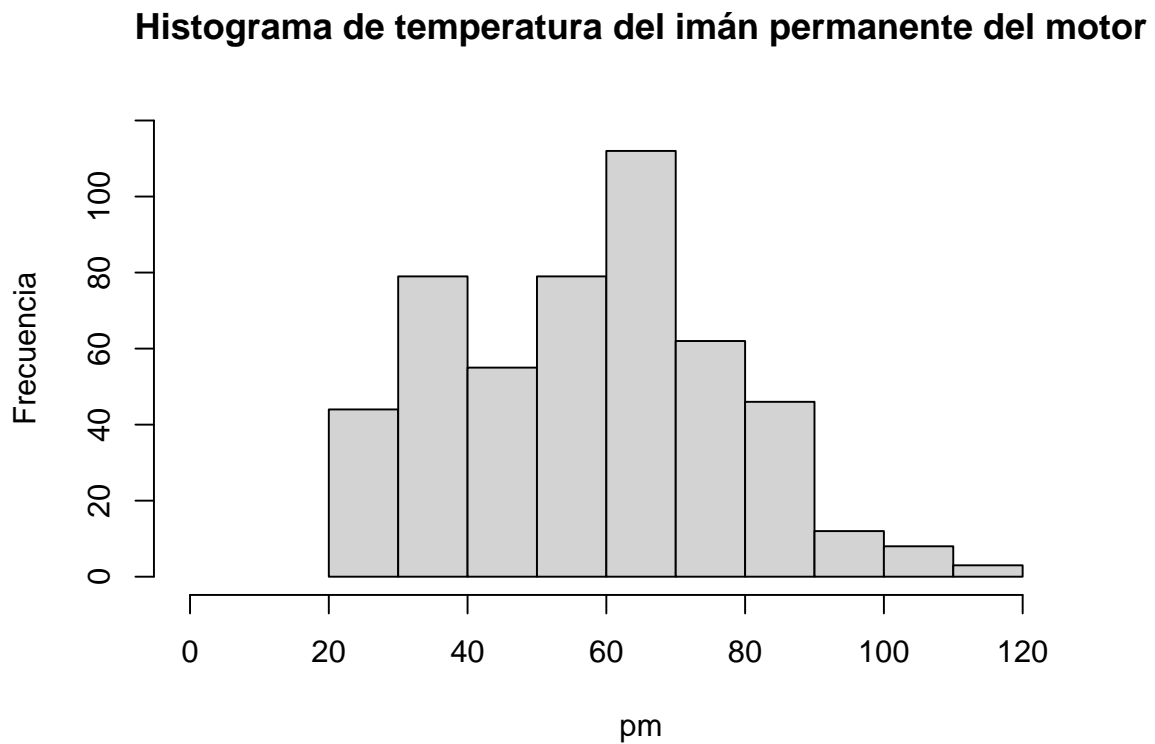


Fig 9: Histograma pm

Observaciones: *Podemos observar que esta variable no sigue una distribución normal.*

Histograma ambient (Temperatura ambiente durante la medición)

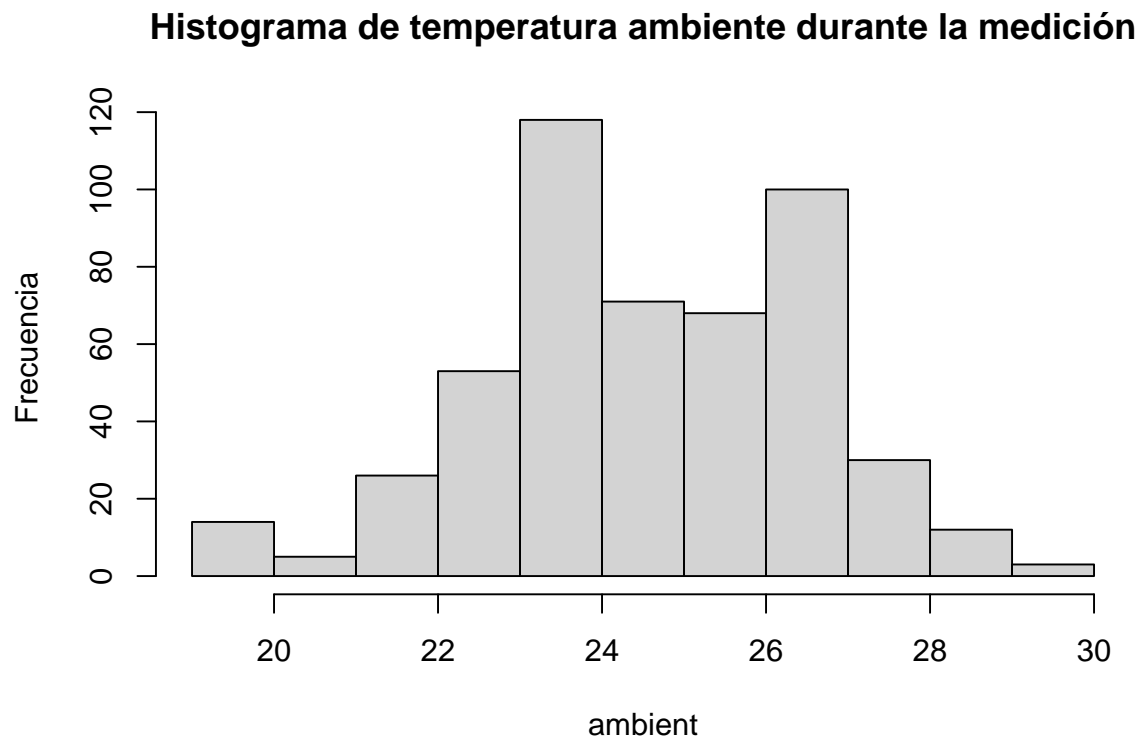


Fig 10: Histograma ambient

Observaciones: *Podemos observar que esta variable podría seguir una distribución de tipo normal dado los extremos.*

Histograma stator_yoke (Temperatura del yugo del estator del motor)

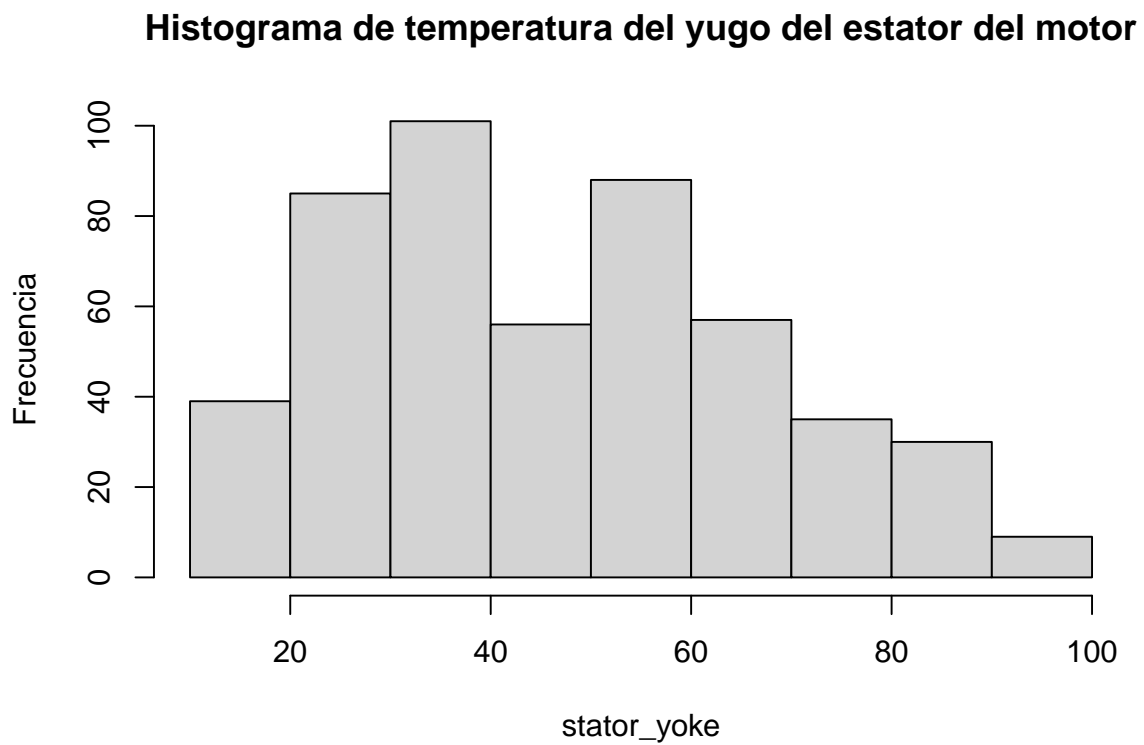


Fig 11: Histograma stator_yoke

Observaciones: Podemos observar que esta variable podría seguir una distribución de tipo normal dado los extremos.

Histograma torque (Par del motor)

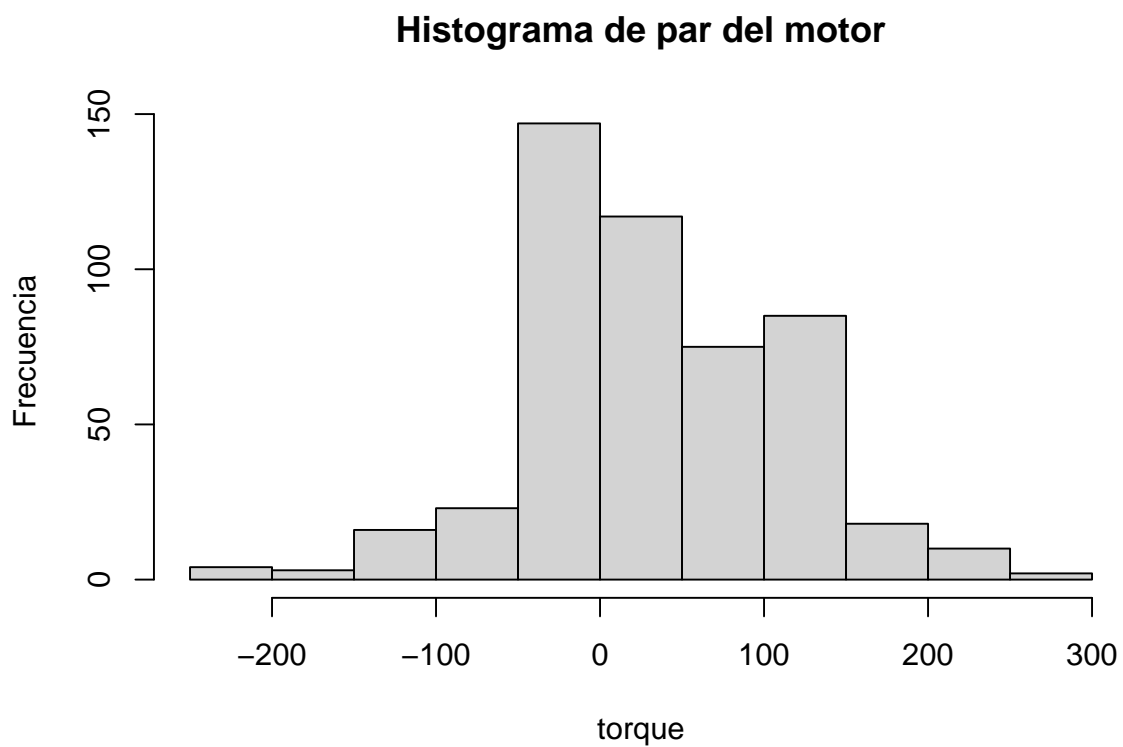


Fig 12: Histograma torque

Observaciones: Podemos observar que esta variable podría seguir una distribución de tipo normal dado los extremos.

Vector de Promedios

Antes de realizar una estandarización de los datos, procederemos a obtener el vector de promedios de cada una de las variables

Tabla 2: Medias del conjunto de datos

Variable	Promedio
u_q	55.6
coolant	35.5
stator_winding	64.7
u_d	-27.5
stator_tooth	55.7
motor_speed	2254.6
i_d	-69.5
i_q	40.3
pm	57.8
stator_yoke	47.2
ambient	24.5
torque	33.4

Observaciones: Se observa que la mayoría de los promedios de los datos se encuentran en un rango similar a excepción de la variable **motor_speed** la cual hace sentido ya que es la velocidad del motor.

Matriz de Correlaciones

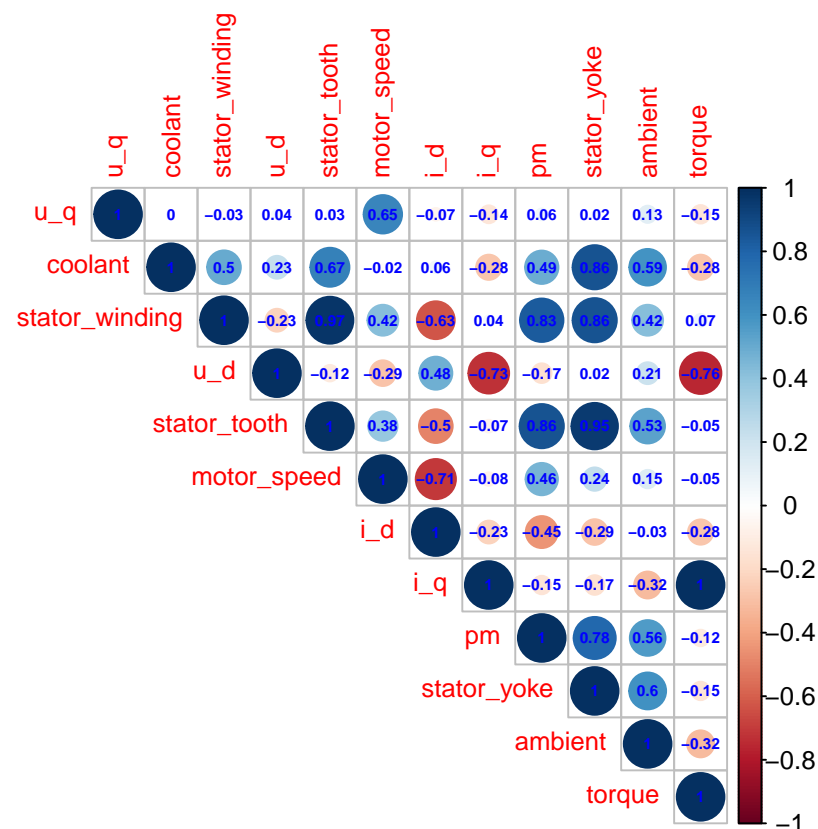


Fig 13: Plot correlaciones

Observaciones: Se observa fuertes correlaciones tanto positivas como negativas. Las correlaciones mas notables mostradas en la gráfica son:

Tabla 3: Variables con mayor correlacion

Relacion	Coeficiente
coolant - stator_tooth	0.67
stator_winding - stator_tooth	0.97
stator_winding - i_d	-0.63
u_q - motor_speed	0.62
u_d - i_q	-0.73
motor_speed - i_d	-0.71
stator_tooth - pm	0.86
stator_winding - pm	0.83
coolant - stator_yoke	0.86
stator_winding - stator_yoke	0.86
stator_tooth - stator_yoke	0.95
pm - stator_yoke	0.78
coolant - ambient	0.59
stator_tooth - stator_yoke	0.95
pm - ambient	0.56
u_d - torque	-0.76
i_q - torque	1

Observaciones: Se observa una cantidad de fuertes correlaciones arriba de 0.70, tanto negativas como positivas. Una de las correlaciones más notorias es una correlación perfecta entre la variable **i_q** y **torque** las cual es de 1.

Prueba de Normalidad

Para esta sección haremos uso de la librería en R **nortest**, la cual aplicaremos para nuestra prueba de normalidad tanto univariada como multivariada. Aunque nuestros datos provienen de una muestra grande, la cual es considerada como $n-p > 40$ la cual se igual a $500 - 13 > 50$, se concluye muestra grande. Se aplicara de igual manera la prueba de normal multivariada.

The figure is a scatter plot with the following characteristics:

- X-axis:** Labeled "Squared Mahalanobis Distance", ranging from 0 to 150 with major ticks at 0, 50, 100, and 150.
- Y-axis:** Labeled "Chi-Square Quantile", ranging from 0 to 30 with major ticks at 5, 10, 15, 20, 25, and 30.
- Theoretical Line:** A solid black line representing the theoretical Chi-Square distribution, starting at (0,0) and curving upwards.
- Data Points:** Approximately 40 black dots representing the observed data. They follow the theoretical line closely for distances up to about 45, after which they curve upwards more steeply than the theoretical line, reaching a quantile of approximately 35 at a distance of 155.

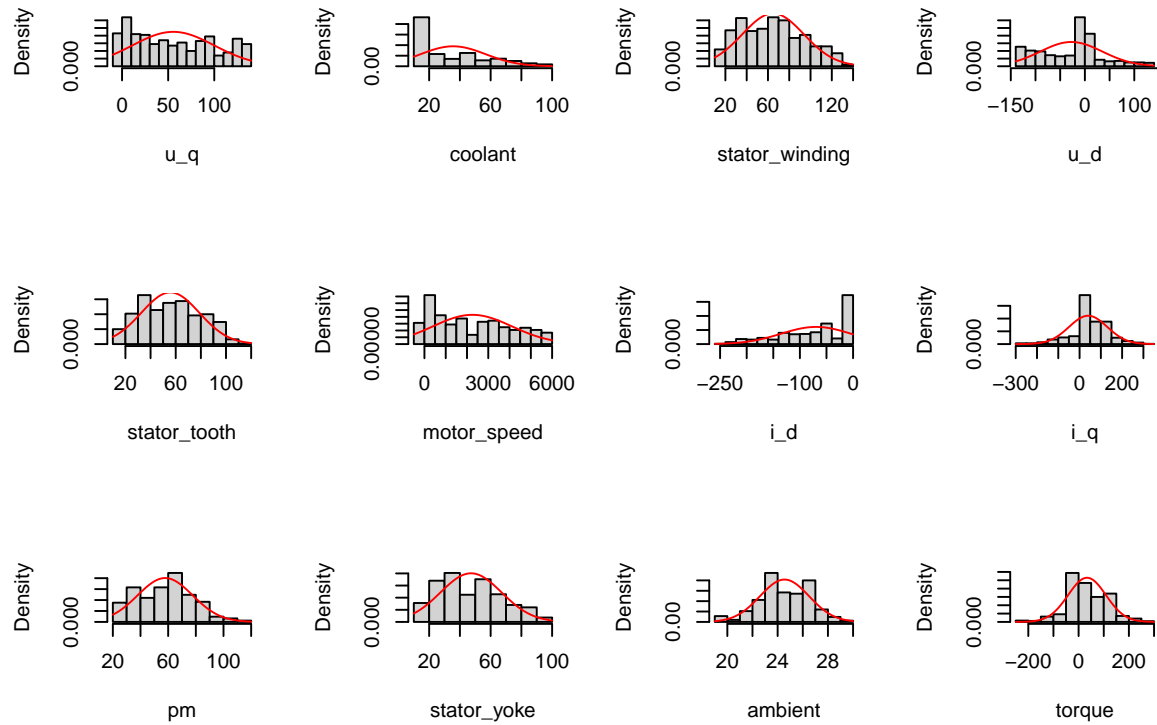
Chi-Square Quantile

Robust Squared Mahalanobis Distance

Outliers (n=244)

Non-outliers (n=256)

Line: 26.962



```
## $multivariateNormality
```

```
##      Test      H      p value MVN
```

```
## 1 Royston 338.9202 1.517446e-69 NO
```

```
##
```

```
## $univariateNormality
```

```
##      Test      Variable Statistic      p value Normality
```

```
## 1 Cramer-von Mises      u_q      1.7079 <0.001      NO
```

```
## 2 Cramer-von Mises      coolant      6.5682 <0.001      NO
```

```
## 3 Cramer-von Mises stator_winding      0.4716 <0.001      NO
```

```
## 4 Cramer-von Mises      u_d      2.5217 <0.001      NO
```

```
## 5 Cramer-von Mises stator_tooth      0.6004 <0.001      NO
```

```
## 6 Cramer-von Mises motor_speed      1.6783 <0.001      NO
```

```
## 7 Cramer-von Mises      i_d      2.6892 <0.001      NO
```

```
## 8 Cramer-von Mises      i_q      1.6388 <0.001      NO
```

```
## 9 Cramer-von Mises      pm      0.3527 1e-04      NO
```

```
## 10 Cramer-von Mises stator_yoke      0.9116 <0.001      NO
```

```
## 11 Cramer-von Mises      ambient      0.5591 <0.001      NO
```

```
## 12 Cramer-von Mises      torque      1.6742 <0.001      NO
```

```
##
```

```
## $Descriptives
```

```
##      n      Mean      Std.Dev      Median      Min      Max
```

```
## u_q      500 55.56100 44.359436 49.790200 -2.29622857 131.995239
```

```
## coolant      500 35.59623 21.101023 26.929087 15.43173929 92.645082
```

```
## stator_winding 500 64.79923 28.914310 63.992186 19.46124458 131.490416
```

```
## u_d      500 -27.51719 62.915778 -7.905764 -130.91442871 130.416699
```

```
## stator_tooth      500 55.73174 23.050638 54.164327 18.89172935 110.369407
```

```
## motor_speed      500 2254.60852 1855.546786 1999.976440 -0.01265136 5999.957031
```

```
## i_d      500 -69.56402 65.385654 -46.203419 -240.66177368 -1.994301
```

```
## i_q      500 40.32900 89.740666 21.085450 -269.57560348 301.703308
```

```
## pm      500 57.80232 19.891348 58.957228 21.54023552 112.971924
```

```
## stator_yoke      500 47.24290 19.895538 45.475449 18.80884743 99.868602
```


## ambient	500	24.54791	1.954389	24.703144	19.34749413	29.549906
## torque	500	33.48011	75.231563	17.894003	-225.55560853	258.307434
##		25th	75th	Skew	Kurtosis	
## u_q		13.2850962	94.3078025	0.2974538	-1.2696476	
## coolant		18.6867557	49.7834830	1.0342137	-0.1116668	
## stator_winding		39.4519179	86.9731099	0.2857520	-0.8505268	
## u_d		-84.3779923	0.8870415	0.1316162	-0.3471005	
## stator_tooth		35.5898691	72.0191218	0.2030437	-1.0135644	
## motor_speed		499.9754379	3740.9020605	0.3643730	-1.1351510	
## i_d		-118.7993946	-3.4489692	-0.7005720	-0.6250497	
## i_q		1.0961654	109.0134249	-0.1520373	0.8415012	
## pm		40.7730618	70.8915024	0.1683998	-0.6501505	
## stator_yoke		30.0590935	60.6632499	0.4716363	-0.6746356	
## ambient		23.1755805	26.2644056	-0.1541419	-0.3213925	
## torque		-0.2358166	95.6061493	-0.1431499	0.8663066	

Fig 14: QQplot Fig 15: Outliers

Establecemos las hipótesis a considerar para normal multivariada y univariada.

Ho: Los datos provienen de una normal multivariada H1: Los datos no provienen de una normal multivariada.

Ho: La variable proviene de una distribución normal, H1: La variable sigue otro tipo de distribución.

Observaciones: Con respecto a la prueba de normal multivariada se establece que se rechaza Ho si p-valor (**0.00000000157**) es menor a alfa 0.05 por lo tanto se rechaza Ho, los datos **no provienen de una distribución normal multivariada**. Para las pruebas de normal univariada se establece que se rechaza Ho si p-valor es menor a alfa 0.05, Se rechaza Ho para todas las variables, **ninguna de las variables cumplió la normal univariada**. Por otra parte la grafica de outliers muestra la mayoría de los puntos como **outliers los cuales fueron 244/500** observaciones los cuales son casi la mitad de las observaciones ,con respecto a la gráfica de QQplot se observa que las observaciones al inicio se ajustaron a la línea de la normal pero la mayoría de las observaciones no pudieron ajustarse a la línea.

Análisis de Factores

El Análisis Factorial es, por tanto, una técnica de reducción de la dimensionalidad de los datos. Su propósito último consiste en buscar el número mínimo de dimensiones capaces de explicar el máximo de información contenida en los datos.

Para desarrollar el análisis de factores se realizaran pasos previos tales como estandarizar los datos, verificar si los datos cumplen la **normal multivariada**, revisar la **matriz de correlaciones** y realizar **supuestos e hipótesis**.

Paso 1: Carga de Datos

```
## # A tibble: 1,330,816 x 12
##       u_q coolant stator_win~1      u_d stato~2 motor_~3      i_d      i_q      pm
##       <dbl>   <dbl>         <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 -0.451    18.8         19.1 -0.350    18.3  2.87e-3  4.42e-3  3.28e-4  24.6
## 2 -0.326    18.8         19.1 -0.306    18.3  2.57e-4  6.06e-4 -7.85e-4  24.5
## 3 -0.441    18.8         19.1 -0.373    18.3  2.35e-3  1.29e-3  3.86e-4  24.5
## 4 -0.327    18.8         19.1 -0.316    18.3  6.10e-3  2.56e-5  2.05e-3  24.6
## 5 -0.471    18.9         19.1 -0.332    18.3  3.13e-3 -6.43e-2  3.72e-2  24.6
## 6 -0.539    18.9         19.1  0.00915   18.3  9.64e-3 -6.14e-1  3.37e-1  24.6
## 7 -0.653    18.9         19.1  0.239     18.3  1.34e-3 -1.01e+0  5.54e-1  24.6
## 8 -0.758    19.0         19.1  0.395     18.3  1.42e-3 -1.29e+0  7.06e-1  24.6
## 9 -0.727    19.0         19.1  0.547     18.3  5.77e-4 -1.49e+0  8.17e-1  24.6
## 10 -0.874   19.0         19.1  0.579     18.3 -1.25e-3 -1.63e+0  8.98e-1  24.6
```

```
## # ... with 1,330,806 more rows, 3 more variables: stator_yoke <dbl>,
## #   ambient <dbl>, torque <dbl>, and abbreviated variable names
## #   1: stator_winding, 2: stator_tooth, 3: motor_speed
```

Paso 2: Estandarizar datos

```
##           u_q      coolant stator_winding      u_d stator_tooth motor_speed
## [1,] -1.27834410 -0.8311962  -1.43987472  0.4603935  -1.4450026  -1.2150662
## [2,] -1.29998946  0.3778941  -0.40021908  0.4642469  -0.2280093  -1.2150626
## [3,]  0.68739232  1.3867850   1.38311958 -1.1322555   1.5957661   1.4795303
## [4,]  0.08919665 -0.5129853   0.08196488 -1.1405594  -0.1277971  -0.2900407
## [5,]  0.93365153 -0.8302683  -1.35853668  0.4270964  -1.2968917  -0.1372255
## [6,] -0.94387184 -0.8088649  -0.87802889  0.2524234  -1.0398530  -1.0803338
##           i_d      i_q      pm stator_yoke      ambient      torque
## [1,]  1.0333025 -0.4371630 -0.09350299 -1.37623227  0.7425053 -0.3724405
## [2,]  1.0333128 -0.4371546  0.71427036 -0.04505362  0.9194281 -0.4450275
## [3,] -0.9582591  0.1539443  1.71937546  1.60801206  0.9222913  0.1856948
## [4,] -0.2317817  1.7182056 -0.41674479 -0.39157610 -0.7370514  1.6244578
## [5,]  1.0332965 -0.4371743 -1.20503791 -1.20390935 -0.3008730 -0.4487295
## [6,]  0.3984230  1.0283933 -1.35716354 -1.04509760 -0.8976195  0.9337019
```

Paso 3: Revisar de cumplimiento de normal multivariada

Para este cumplimiento de normal multivariada creamos nuestras hipótesis:

$$H_0 : \mu_1 = \mu_2 = \mu_3 \dots \mu_k$$

$$H_1 : \mu_1 \neq \mu_2 \dots \neq \mu_k$$

```
##           Test      HZ p value MVN
## 1 Henze-Zirkler 4.782402      0 NO
```

Para el cumplimiento de normal univariada creamos de igual manera nuestras hipótesis:

H₀ : los datos provienen de una distribución normal.

H₁ : los datos provienen de otra distribución.

```
##           Test      Variable Statistic  p value Normality
## 1 Anderson-Darling      u_q      12.0822 <0.001      NO
## 2 Anderson-Darling    coolant      37.8195 <0.001      NO
## 3 Anderson-Darling stator_winding    3.6821 <0.001      NO
## 4 Anderson-Darling      u_d      13.2541 <0.001      NO
## 5 Anderson-Darling stator_tooth    4.5108 <0.001      NO
## 6 Anderson-Darling motor_speed     11.8267 <0.001      NO
## 7 Anderson-Darling      i_d      17.9452 <0.001      NO
## 8 Anderson-Darling      i_q       8.2897 <0.001      NO
## 9 Anderson-Darling      pm       2.6402 <0.001      NO
## 10 Anderson-Darling stator_yoke     6.2023 <0.001      NO
## 11 Anderson-Darling    ambient     3.2206 <0.001      NO
## 12 Anderson-Darling    torque     8.5511 <0.001      NO
```

Observaciones: Se observa que no se cumplió con la prueba de normal multivariada dado su *p-valor* es **0** menor a alfa **0.05**, se rechaza **H₀** los datos **no provienen de una normal multivariada**, con respecto a las pruebas de **normalidad univariada** se observa que **ninguna variable** cumple con normalidad

dados sus *p-valores* cercanos al cero y menores a alfa **0.05** por lo tanto los datos siguen otro tipo de distribución.

Matriz de Correlaciones

Análisis incluido en la sección **Análisis exploratorio** - *subsección: Matriz de Correlaciones*

Paso 4: Prueba de esfericidad

Para esta prueba se usara la prueba de esfericidad de Bartlett la cual sirve para identificar si la correlación entre pares de variables es cero o no.

Definimos nuestras hipótesis

Ho: La correlación entre cada par de variables es cero H1: La correlación entre cada par de variable diferente de cero

```
## $chisq
## [1] 2300.478
##
## $p.value
## [1] 0
##
## $df
## [1] 66
```

Observaciones: Dado que el *p_valor* es menor a alfa **0.05** , se rechaza Ho por lo tanto las correlaciones son diferente de 0.

Paso 5: Determinar numero de factores

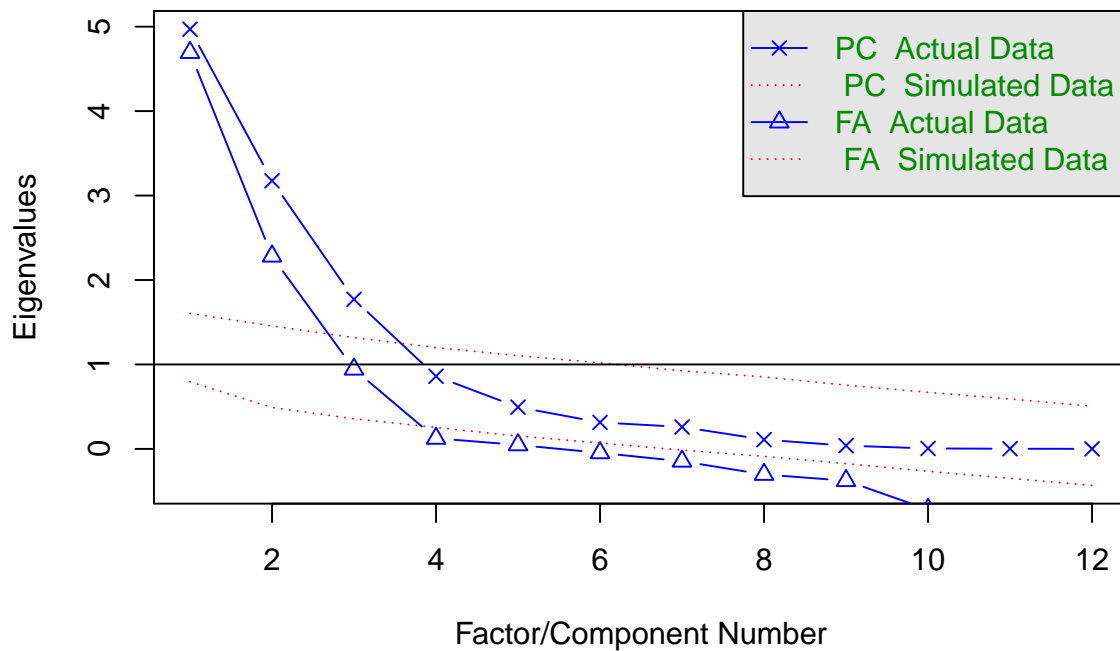
Prueba de PCA Para determinar el número de factores, procederemos a realizar un Análisis de Componentes Principales (PCA), el cual nos sugerirá el número de factores a considerar.

```
## Importance of components:
##              PC1    PC2    PC3    PC4    PC5    PC6    PC7
## Standard deviation  2.2293 1.7819 1.3309 0.92737 0.7031 0.56018 0.50943
## Proportion of Variance 0.4142 0.2646 0.1476 0.07167 0.0412 0.02615 0.02163
## Cumulative Proportion 0.4142 0.6787 0.8263 0.89801 0.9392 0.96536 0.98698
##              PC8    PC9    PC10    PC11    PC12
## Standard deviation  0.33131 0.19658 0.07310 0.04296 0.02463
## Proportion of Variance 0.00915 0.00322 0.00045 0.00015 0.00005
## Cumulative Proportion 0.99613 0.99935 0.99980 0.99995 1.00000
```

Observaciones: Se puede observar que el número de factores óptimo esta entre 3 y 4, dada la varianza acumulada que contienen.

Grafica de Codo Creamos nuestra gráfica de codo la cual no sugiere de igual manera elegir entre 3 y 4 factores.

Parallel Analysis Scree Plots



Parallel analysis suggests that the number of factors = 3 and the number of components = 3

Fig 16: Gráfica de Codo

Observaciones: Se puede observar que al elegir 3 factores obtenemos **82%** de la varianza explicada, la cual es un buen porcentaje, Procedemos a usar el algoritmo ahora rotando los ejes usando el método de 'varimax'.

Factores usando Varimax

```
## Principal Components Analysis
## Call: principal(r = R, nfactors = 3, rotate = "varimax")
## Standardized loadings (pattern matrix) based upon correlation matrix
##
```

	RC1	RC2	RC3	h2	u2	com
## u_q	-0.09	0.23	0.77	0.65	0.349	1.2
## coolant	0.78	0.31	-0.22	0.76	0.244	1.5
## stator_winding	0.90	-0.24	0.21	0.92	0.084	1.3
## u_d	-0.04	0.88	-0.20	0.81	0.193	1.1
## stator_tooth	0.97	-0.09	0.16	0.97	0.034	1.1
## motor_speed	0.24	-0.10	0.95	0.97	0.035	1.1
## i_d	-0.37	0.50	-0.59	0.74	0.265	2.7
## i_q	-0.13	-0.92	-0.14	0.89	0.112	1.1
## pm	0.85	-0.04	0.28	0.81	0.192	1.2
## stator_yoke	0.98	0.07	0.01	0.96	0.044	1.0
## ambient	0.66	0.35	0.03	0.55	0.448	1.5
## torque	-0.11	-0.94	-0.11	0.91	0.086	1.1

```
##
##
```

	RC1	RC2	RC3
## SS loadings	4.70	3.10	2.11
## Proportion Var	0.39	0.26	0.18

```
## Cumulative Var      0.39 0.65 0.83
## Proportion Explained 0.47 0.31 0.21
## Cumulative Proportion 0.47 0.79 1.00
##
## Mean item complexity = 1.3
## Test of the hypothesis that 3 components are sufficient.
##
## The root mean square of the residuals (RMSR) is 0.07
##
## Fit based upon off diagonal values = 0.98
```

Observaciones: Se observa una varianza acumulada del **83 %**, con respecto a los **residuales RMSR** se observa un valor muy bajo de **0.07** cercano a cero. Con respecto a las cargas elegidas estas muestran comunalidades (

$$h_2$$

) altas y la varianza no explicada

$$u_2$$

es muy baja. También observamos con el método de *varimax* de una manera muy clara los **variables dominantes para cada factor** los cuales son:

- **Factor 1 :** stator_winding, stator_tooth, pm, stator_yoke, coolant, ambient
- **Factor 2 :** u_d,i_q,torque
- **Factor 3 :** u_q,motor_speed,i_d,

Paso 6: Representación gráfica

Representación grafica de cada uno de las variables.

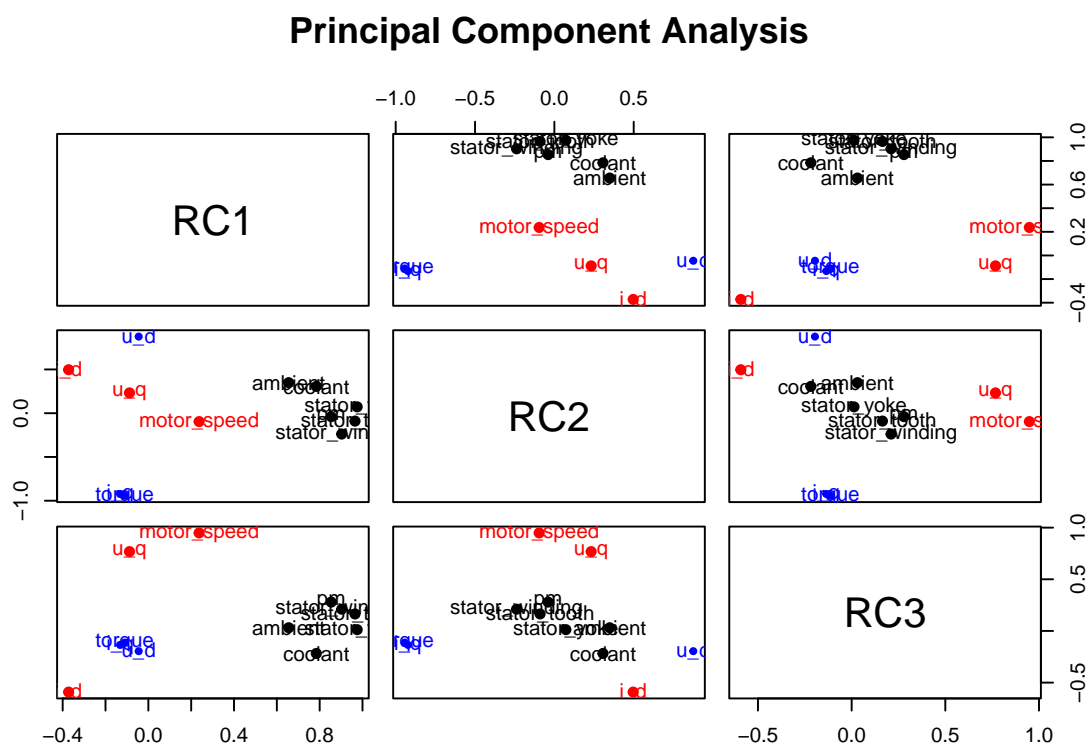


Fig 17: Grafica Factores

Observaciones: Se puede observar una agrupación muy notoria en las variables “stator_winding, stator_tooth, pm, stator_yoke, coolant, ambient”, mientras que motor_speed y u_q están cercanas entre ellas, también se observa que las variable torque y u_d están muy cercanas, la única variable que está muy alejada de los grupos antes mencionados es la variable i_d.

Conclusiones

Se concluye que aunque no se cumplieron los supuesto de normal multivariada dadas las pruebas de hipótesis, se obtuvo una varianza acumulada de 82% usando 3 factores con lo cual se redujo la dimensión de variables de 12 variables a solo 3, Por otra estos factores mostraron **comunalidades muy altas y varianza no explicada muy baja**, Con respecto a las variables dominantes de cada factor estas quedaron de la siguiente forma:

- **Factor 1** : stator_winding, stator_tooth, pm, stator_yoke, coolant, ambient
- **Factor 2** : u_d,i_q,torque
- **Factor 3** : u_q,motor_speed,i_d

Nota: Dados las variables dominantes para cada factor daremos un nombre que haga sentido a los factores.

- **Factor 1** : ‘Componentes del motor’
- **Factor 2** : ‘Variables eléctricas y de torque’
- **Factor 3** : ‘Variables de velocidad y corriente directa’

Análisis de Componentes Principales

El análisis de componentes principales (PCA) es un método estadístico que sirve para simplificar y resumir un conjunto de datos con muchas variables. En vez de usar todas las variables originales, el PCA encuentra mezclas de estas variables que capturan la mayor parte de la variación en los datos. Estas mezclas se llaman “componentes principales” y se usan para describir el conjunto de datos de una manera más sencilla y comprensible.

Piensa que tienes muchos datos distintos sobre un motor eléctrico, como la temperatura, el torque, la velocidad, etc. Puede ser complicado saber cómo todos estos datos se relacionan entre sí y qué información es la más relevante. Si el análisis de componentes principales descubre que la temperatura y el torque están muy relacionados, entonces puede juntarlos en un nuevo componente principal que abarque ambas variables a la vez. Así, se puede disminuir el número de variables y hacer que los datos sean más sencillos de entender.

Paso 1: Carga de datos y preprocesamiento

En este primer paso se procede a la carga de los datos, los cuales deben ser preprocesados para asegurar que el análisis posterior sea correcto. En el código se utiliza la función read.csv para leer el archivo de datos measures_v2.csv, y se utiliza la librería dplyr para eliminar la columna profile_id, que contiene identificadores únicos para cada observación y no aporta información relevante para el análisis.

```
##           u_q  coolant stator_winding           u_d stator_tooth  motor_speed
## 1 -0.4506815 18.80517      19.08667 -0.350054592      18.29322 0.0028655678
## 2 -0.3257370 18.81857      19.09239 -0.305803001      18.29481 0.0002567817
## 3 -0.4408640 18.82877      19.08938 -0.372502625      18.29409 0.0023549714
## 4 -0.3270257 18.83557      19.08303 -0.316198707      18.29254 0.0061046658
## 5 -0.4711501 18.85703      19.08253 -0.332272142      18.29143 0.0031328229
## 6 -0.5389726 18.90155      19.07711  0.009147473      18.29063 0.0096361237
##           i_d           i_q           pm stator_yoke  ambient  torque
## 1 4.419137e-03 0.0003281022 24.55421      18.31655 19.85069 0.1871008
```

```
## 2  6.058724e-04 -0.0007853527 24.53808    18.31495 19.85067 0.2454175
## 3  1.289587e-03  0.0003864682 24.54469    18.32631 19.85066 0.1766153
## 4  2.558433e-05  0.0020456610 24.55402    18.33083 19.85065 0.2383027
## 5 -6.431678e-02  0.0371837765 24.56540    18.32666 19.85064 0.2081967
## 6 -6.136352e-01  0.3367473483 24.57360    18.32386 19.85063 0.4762178
```

Paso 2: Estandarización de datos

Antes de realizar el análisis de componentes principales, es necesario estandarizar los datos. En este paso se utiliza la función `fn_std` para estandarizar los datos de cada variable, es decir, se resta la media y se divide por la desviación estándar. Luego se aplica esta función a cada columna del dataframe `df` utilizando la función `apply`, y se muestran los primeros registros de los datos estandarizados.

```
##          u_q      coolant stator_winding      u_d stator_tooth motor_speed
## [1,] -1.238978 -0.7998120   -1.648158 0.3928195   -1.681116   -1.184127
## [2,] -1.236150 -0.7991969   -1.647958 0.3935209   -1.681047   -1.184129
## [3,] -1.238756 -0.7987288   -1.648063 0.3924637   -1.681078   -1.184127
## [4,] -1.236179 -0.7984168   -1.648285 0.3933561   -1.681145   -1.184125
## [5,] -1.239442 -0.7974315   -1.648302 0.3931013   -1.681194   -1.184127
## [6,] -1.240977 -0.7953882   -1.648491 0.3985128   -1.681229   -1.184124
##          i_d      i_q      pm stator_yoke      ambient      torque
## [1,] 1.058337 -0.4058548 -1.786836   -1.494243 -2.443385 -0.4008379
## [2,] 1.058278 -0.4058668 -1.787685   -1.494323 -2.443395 -0.4000818
## [3,] 1.058289 -0.4058541 -1.787337   -1.493755 -2.443403 -0.4009738
## [4,] 1.058269 -0.4058361 -1.786846   -1.493528 -2.443408 -0.4001741
## [5,] 1.057278 -0.4054549 -1.786248   -1.493737 -2.443412 -0.4005644
## [6,] 1.048818 -0.4022052 -1.785816   -1.493877 -2.443415 -0.3970897
```

De igual manera se validan las dimensiones del dataset con `dim()`.

Tabla 4: Tabla de dimensiones del dataset

Filas	Columnas
1330816	12

Paso 3: Análisis de componentes principales

El análisis de componentes principales (PCA) se utiliza para identificar patrones en los datos y reducir la dimensionalidad de los mismos. En este paso se utiliza la función `prcomp` para realizar el análisis de componentes principales sobre los datos estandarizados.

```
## Standard deviations (1, ..., p=12):
## [1] 2.20580098 1.77382168 1.31839690 0.94318021 0.76770930 0.58383387
## [7] 0.52207927 0.33680085 0.18901079 0.07615769 0.04278316 0.02535302
##
## Rotation (n x k) = (12 x 12):
##          PC1          PC2          PC3          PC4          PC5
## u_q      0.10700072  0.01661259  0.574460657 -0.595496339  0.2193584376
## coolant  0.31078520  0.21312209 -0.288548744 -0.309999643  0.3336760048
## stator_winding 0.41588420 -0.12735453 -0.094115390  0.200596971  0.1205939851
## u_d      -0.05306894  0.49366940  0.024871306  0.128221868  0.1853026530
## stator_tooth  0.43832602 -0.03926350 -0.122777788  0.063301863  0.1526011702
## motor_speed  0.24295909 -0.15374050  0.588758490 -0.007240699  0.0171154510
## i_d      -0.23488550  0.32097641 -0.257817655 -0.454883976  0.0721479240
## i_q      -0.06611979 -0.49905940 -0.226244621 -0.286189478  0.0116207365
```

## pm	0.39880295	-0.01336623	-0.003557004	0.112557069	-0.2957437807
## stator_yoke	0.42185478	0.06471435	-0.210672036	-0.094462485	0.2338878077
## ambient	0.25586599	0.22701419	-0.069899226	-0.340256158	-0.7877650919
## torque	-0.05489022	-0.51108550	-0.220475643	-0.249708688	-0.0002577344
##	PC6	PC7	PC8	PC9	PC10
## u_q	-0.001712885	-0.25182557	-0.15571150	0.40826759	-0.034208524
## coolant	-0.066696030	0.52618574	0.33932067	0.10517559	-0.384147757
## stator_winding	-0.094112782	-0.28808332	-0.41281209	-0.12599407	-0.593847224
## u_d	-0.540101985	-0.49343068	0.39049596	-0.11149535	0.003287108
## stator_tooth	-0.001323098	-0.17394160	-0.25485258	-0.12150953	0.230889314
## motor_speed	-0.008406877	0.21226641	0.24515612	-0.68091441	0.001735467
## i_d	0.471842392	-0.25741885	-0.12402040	-0.50159335	-0.077394267
## i_q	-0.208755737	-0.20460606	0.22349260	-0.10223947	0.032535497
## pm	0.531110752	-0.34274389	0.53864350	0.21520428	-0.063674866
## stator_yoke	-0.013369836	0.03804004	-0.06317565	-0.02181754	0.658395937
## ambient	-0.333783659	0.07692902	-0.14410970	-0.06031060	-0.004827930
## torque	-0.187151797	-0.15801046	0.18010427	-0.05537057	-0.025896349
##	PC11	PC12			
## u_q	0.002549219	3.396984e-03			
## coolant	-0.021003495	1.084474e-01			
## stator_winding	-0.008434708	-3.410545e-01			
## u_d	0.042492461	1.260929e-02			
## stator_tooth	-0.041997105	7.738957e-01			
## motor_speed	0.020184334	-3.269149e-02			
## i_d	0.030348780	-2.051029e-02			
## i_q	-0.681896572	-3.815568e-02			
## pm	0.002829883	-2.381642e-02			
## stator_yoke	0.060810404	-5.162995e-01			
## ambient	0.002296022	3.480655e-05			
## torque	0.725182231	5.251892e-02			

Estas desviaciones estándar indican cuánto varían los datos proyectados en cada componente principal. A mayor desviación estándar, mayor variación y mayor importancia del componente principal.

Se puede observar que el primer componente principal tiene la mayor desviación estándar (2.2), lo que significa que explica la mayor parte de la variación de los datos originales. El segundo componente principal tiene la segunda mayor desviación estándar (1.77), lo que significa que explica la mayor parte de la variación que queda después de quitar el efecto del primero. Y así sucesivamente con los demás componentes principales.

Se puede ver también que las desviaciones estándar van disminuyendo a medida que aumenta el número de componente principal, lo que significa que cada componente principal explica menos variación que el anterior. Los últimos componentes principales tienen desviaciones estándar muy pequeñas (menores a 0.1), lo que significa que explican muy poca variación de los datos y que podrían ser ignorados sin perder mucha información.

Sobre los resultados de la matriz de rotación y los coeficientes de los componentes principales obtenidos por el PCA, la matriz de rotación es una matriz que contiene los vectores propios de los datos, es decir, las direcciones en las que los datos varían más. Cada fila de la matriz representa un vector propio y cada columna representa una coordenada del vector. Los coeficientes de los componentes principales son los números que indican cómo se combinan las variables originales para formar cada componente principal. Cada fila de la tabla representa una variable original y cada columna representa un componente principal.

Los números pueden ser positivos o negativos. Los números positivos indican que la variable original y el primer componente principal se mueven en la misma dirección. Por ejemplo, si la variable coolant aumenta, el primer componente principal también aumenta. Los números negativos indican que la variable original y el primer componente principal se mueven en direcciones opuestas. Por ejemplo, si la variable u_d aumenta, el primer componente principal disminuye.

Los números grandes indican que la variable original tiene mucha influencia en el primer componente principal. Por ejemplo, la variable stator_yoke tiene el número más grande (0.42185478), lo que significa

que es la que más contribuye a la diferencia entre los datos. Los números pequeños indican que la variable original tiene poca influencia en el primer componente principal. Por ejemplo, la variable torque tiene el número más pequeño (-0.05489022), lo que significa que es la que menos contribuye a la diferencia entre los datos.

El primer componente principal es importante porque explica la mayor parte de la diferencia entre los datos. El doceavo componente principal es el último que se puede hacer con las variables originales. Es el que explica la menor parte de la diferencia entre los datos

Paso 4: Selección de componentes principales

En este paso se seleccionan los componentes principales que explican más del 80 % de la varianza en los datos. Primero se utiliza la función cumsum para calcular la suma acumulativa de las varianzas explicadas por cada componente principal, y luego se determina el número de componentes necesarios para alcanzar el 80 % de la varianza explicada. En este caso, se obtiene que se necesitan 3 componentes principales.

Tabla 5: Tabla de varianza acumulada explicada por cada componente principal

Componente	Varianza_acumulada
1	0.4054632
2	0.6676668
3	0.8125143
4	0.8866467
5	0.9357615
6	0.9641667
7	0.9868806
8	0.9963335
9	0.9993106
10	0.9997939
11	0.9999464
12	1.0000000

Paso 5: Visualización de componentes principales

Finalmente, se utiliza la función `pca1$rotation` para obtener la matriz de rotación, que contiene las cargas de cada variable en cada componente principal, y se muestran los primeros 3 componentes principales y sus respectivas cargas.

Tabla 6: Tabla de componentes principales

	PC1	PC2	PC3
u_q	0.1070007	0.0166126	0.5744607
coolant	0.3107852	0.2131221	-0.2885487
stator_winding	0.4158842	-0.1273545	-0.0941154
u_d	-0.0530689	0.4936694	0.0248713
stator_tooth	0.4383260	-0.0392635	-0.1227778
motor_speed	0.2429591	-0.1537405	0.5887585
i_d	-0.2348855	0.3209764	-0.2578177
i_q	-0.0661198	-0.4990594	-0.2262446
pm	0.3988029	-0.0133662	-0.0035570
stator_yoke	0.4218548	0.0647144	-0.2106720
ambient	0.2558660	0.2270142	-0.0698992
torque	-0.0548902	-0.5110855	-0.2204756

Los valores en la tabla indican cómo cada variable contribuye a cada una de las tres componentes principales.

El primer componente principal tiene valores grandes en las variables `coolant`, `stator_winding`, `stator_tooth`, `i_d`, `pm` y `stator_yoke`. Esto significa que estas variables contribuyen mucho a la diferencia entre los datos en la dirección del primer componente principal y que están correlacionadas entre sí. Podríamos interpretar este componente principal como un factor relacionado con la temperatura del motor.

El segundo componente principal tiene valores grandes en las variables `u_d`, `i_q` y `torque`. Esto significa que estas variables contribuyen mucho a la diferencia entre los datos en la dirección del segundo componente principal y que están correlacionadas entre sí. Podríamos interpretar este componente principal como un factor relacionado con la potencia del motor.

El tercer componente principal tiene valores grandes en las variables `u_q` y `motor_speed`. Esto significa que estas variables contribuyen mucho a la diferencia entre los datos en la dirección del tercer componente principal y que están correlacionadas entre sí. Podríamos interpretar este componente principal como un factor relacionado con la velocidad del motor.

Conclusiones

En esta sección se detalla las conclusiones de cada una de las secciones principales , tanto del Análisis exploratorio como de los métodos aplicados (Análisis de Factores y PCA)

Análisis exploratorio.

En este análisis se observaron que las variables sus histogramas a manera de sospecha seguían una distribución de tipo normal, otras asemejaban otro tipo de distribución, al realizar la prueba de normalidad tanto univariada como multivariada concluimos que los datos no seguían una distribución normal multivariada y individualmente ninguna de las variables seguía una distribución de tipo normal dados sus p-valores para cada uno de las pruebas. Con respecto a la grafica de QQplot y de outliers se observo que las observaciones para QQplot no se ajustaban a la línea y la parte de los outliers la mitad de las observaciones de nuestro conjunto de datos se clasificaron como outliers.

Análisis de Factores

En esta sección se reviso la matriz de correlaciones la cual se observaron correlación muy altas tanto positivas como negativas también se observaron variables independientes la cual su correlación entre variable fue 0, Con respecto a la prueba de esfericidad se establecieron las hipótesis correspondientes y se aplico la prueba la cual se concluyo que las correlación entre las variables son diferentes de 0 dado su p-valor. Para la parte de el numero optimo de factores se aplico una grafica de codo la cual observamos que el numero optimo de factores se encontraba entre 3 y 4 factores, al aplicar el algoritmo de factores usando varimax para la rotación de vectores, se obtuvo una varianza explicada del **82 %** la cual es un buen porcentaje, con respecto a los residuales RSMR los valores fueron muy bajos cercanos a 0, se mostraron comunialidades alta y varianza explicada muy baja. **Concluimos para este método que pudimos reducir nuestro numero de variables de 12 a solo 3 variables manteniendo un porcentaje muy bueno de varianza explicada, se revisó de acuerdo a la variables dominantes de cada uno de los factores y se dieron nombres que dieran sentido a estas nuevas variables.**

Análisis de Componentes Principales

Los resultados obtenidos con el PCA nos pueden servir para realizar una investigación más avanzada o aplicar modelos no paramétricos o de deep learning al conjunto de datos. Al reducir el número de variables de 12 a 3, podemos disminuir la complejidad y el tiempo de cómputo de los modelos, así como evitar el sobreajuste o la maldición de la dimensionalidad. Además, al tener variables más simples y significativas, podemos facilitar la interpretación y la visualización de los resultados.

Por ejemplo, podríamos usar los tres componentes principales como variables predictoras en un modelo no paramétrico como un árbol de decisión o un bosque aleatorio para predecir el consumo eléctrico del motor en función de las características del mismo. O podríamos usar los tres componentes principales como variables de entrada en un modelo de deep learning como una red neuronal artificial o una red convolucional para clasificar el tipo o el estado del motor según su consumo eléctrico.

Referencias

- Jolliffe, I. T. (2002). Principal component analysis. Springer.
- Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4), 433–459. <https://doi.org/10.1002/wics.101>
- Jun Li and Thangarajah Akilan. Global attention-based encoder-decoder lstm model for temperature prediction of permanent magnet synchronous motors. *IEEE*, 11, 2022.
- Kirgiz, W. (2021). Electric Motor Temperature. Recuperado el 10 de enero de 2023, de <https://www.kaggle.com/datasets/wkirgsn/electric-motor-temperature>
- Wickham, H., & Grolemund, G. (2016). *R for data science: Import, tidy, transform, visualize, and model data*. O'Reilly Media.
- Kassambara, A. (2017). QQ-plots: Quantile-Quantile plots - R Base Graphs. STHDA. <http://sthda.com/english/wiki/qq-plots-quantile-quantile-plots-r-base-graphs>
- Lever, J., Krzywinski, M., & Altman, N. (2017). Principal component analysis. *Nature Methods*, 14(7), 641–642. <https://doi.org/10.1038/nmeth.4346>
- Revelle, W. (2018). How to: Use the psych package for factor analysis and data reduction. <https://cran.r-project.org/web/packages/psychTools/vignettes/factor.pdf>
- Kassambara, A. (2019). Mauchly's test of sphericity in R: The definitive guide. Datanovia. <https://www.datanovia.com/en/lessons/mauchlys-test-of-sphericity-in-r/>
- Statology. (2019). A guide to Bartlett's test of sphericity. <https://www.statology.org/bartletts-test-of-sphericity/>
- Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2020). afex: Analysis of factorial experiments. https://cran.r-project.org/web/packages/afex/vignettes/assumptions_of_ANOVAs.html