PRACTICA DE DATAMINING. SAS.

Pedro García Fernández.

PRIMERA PARTE. REGRESIÓN.

Objetivo: Predecir el peso de un niño al nacer.

Datos:

Weight	Infant Birth Weight	Peso al nacer infantil (Target continua)
Black	Black Mother	Madre de color (1 si, 0 no)
Boy	Baby Boy	Bebé (1 niño, 0 niña)
CigsPerDay	Cigarettes Per Day	Número de Cigarrillos por día
Married	Married Mother	Madre casada (1 si, 0 no)
MomAge	Mother's Age	Edad de la madre (equivale el 0 a tener 25 años)
MomEdLevel	Mother's Education Level	Nivel de educación de la madre (de 0 a 3, siendo 0 nada)
MomSmoke	Smoking Mother	Madre fumadora (1 si, 0 no)
MomWtGain	Mother's Pregnancy Weight Gain	Embarazo de la madre pérdida/aumento de peso
Visit	Prenatal Visit	Visita prenatal (de 0 a 3, siendo el 0 ninguna visita)

1. Análisis exploratorio de los datos

Tras la carga de los datos, mostramos los primeros registros de la tabla.

Primeros registros del dataset

Obs	Weight	Black	Married	Boy	MomAge	MomSmoke	CigsPerDay	MomWtGain	Visit	MomEdLevel
1	4111	0	1	1	-3	0	0	-16	1	0
2	3997	0	1	0	1	0	0	2	3	2
3	3572	0	1	1	0	0	0	-3	3	0
4	1956	0	1	1	-1	0	0	-5	3	2
5	3515	0	1	1	-6	0	0	-20	3	0
6	3757	0	1	0	3	0	0	0	3	2
7	2977	1	0	1	-5	1	5	5	3	0
8	3884	0	0	0	-5	0	0	0	3	2
9	3629	0	1	0	6	0	0	-5	3	0
10	3062	0	1	1	-1	0	0	6	3	2
11	4026	0	1	1	-2	1	4	22	3	1
12	3642	0	1	1	-6	0	0	-1	3	0
13	2296	0	0	1	0	0	0	7	3	0
14	2665	0	0	1	1	1	10	-6	3	1
15	2948	0	1	1	1	0	0	10	3	1

Se puede observar lo que se indicaba en el enunciado, que hay varias variables categóricas, tales como Black, Married, Boy MomSmoke y MomEdLevel.
Visit aunque no es contínua, se considera como categórica, aunque sea un contador de visitas, no conozco el dominio del negocio para establecer dichas categorías, por ejemplo como si fueran dos categorías como de 0 a 1 visitas ha ido poco, y de 2 a 3 visitas ha hecho las revisiones oportunas, simplemente las dejo como mero contador de visitas.

Y claramente se ve lo que se indicaba que los valores a los que está referenciada la edad de la madre parten de 0, ya que hay valores negativos. Para hacerla más entendible la normalizo a edades reales y elimino la variable MomAge, ya que sólo voy a usar la nueva que he creado xMomAge.

Registros de la edad de la madre

Obs	xMomAge
1	22
2	26
3	25
4	24
5	19
6	28
7	20
8	20
9	31
10	24

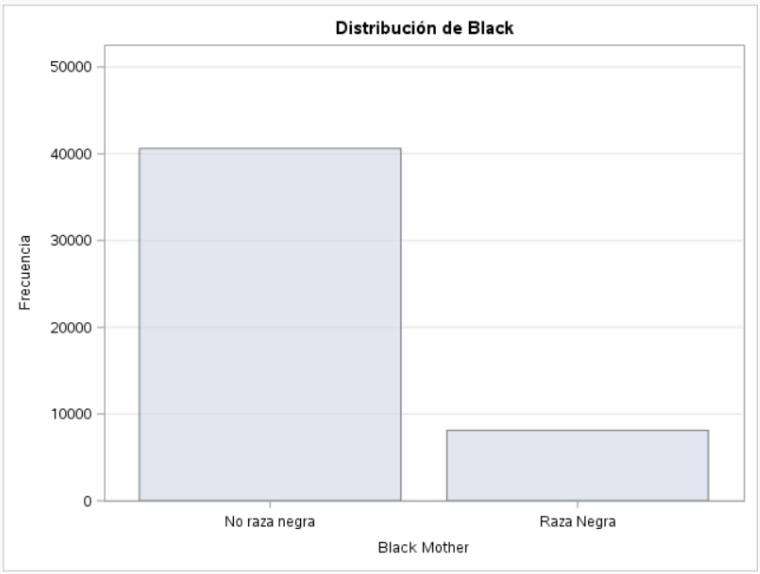
Buscamos si hay registros duplicados y los eliminamos. De los 50000 se descubren unos 1300 duplicados

Nº total de filas: 48734 Nº total de columnas: 10

Muestro a continuación las tablas de frecuencias de las variables categóricas. Previamente he etiquetado los valores de las variables catégoricas para que sean más entendibles.

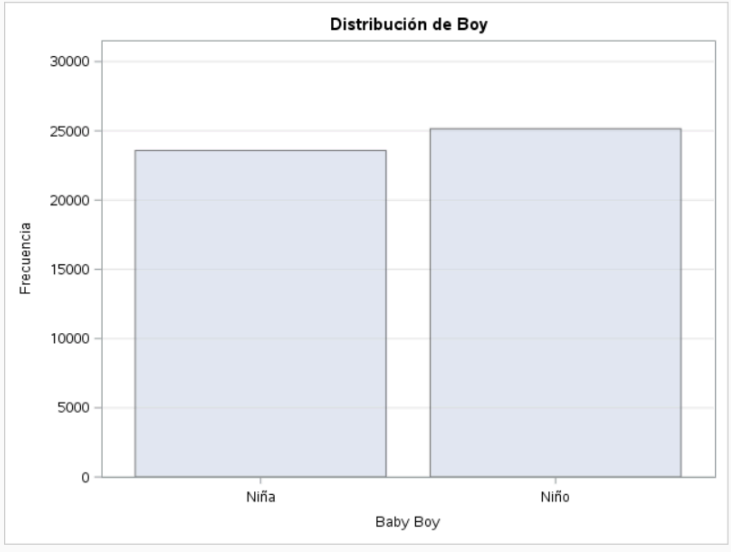
Raza: Hay mayoría de madres que no son de raza negra, un 83%

Black Mother		
Black	Frecuencia	Porcentaje
No raza negra	40620	83.35
Raza Negra	8114	16.65



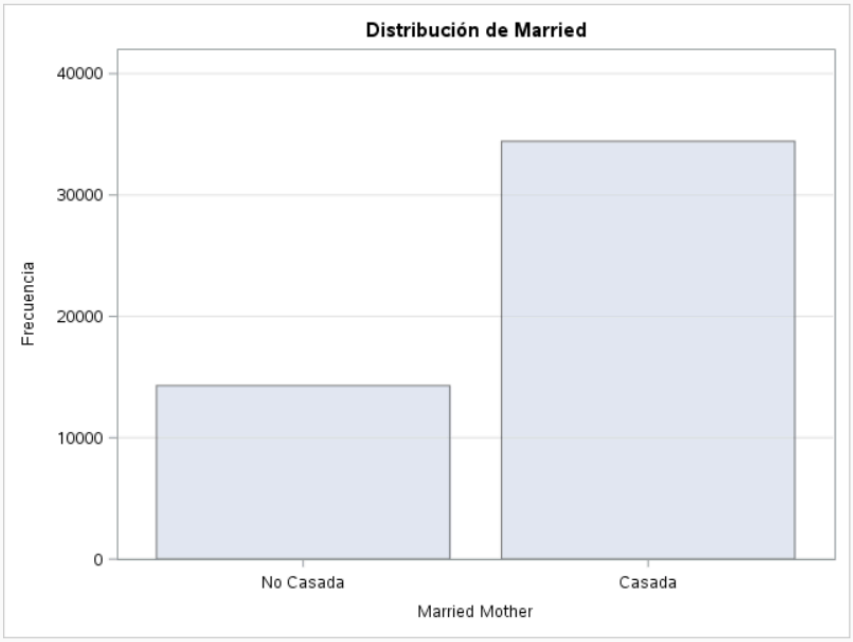
Sexo: Prácticamente, misma distribución de niños y niñas.

Baby Boy		
Boy	Frecuencia	Porcentaje
Niña	23580	48.39
Niño	25154	51.61



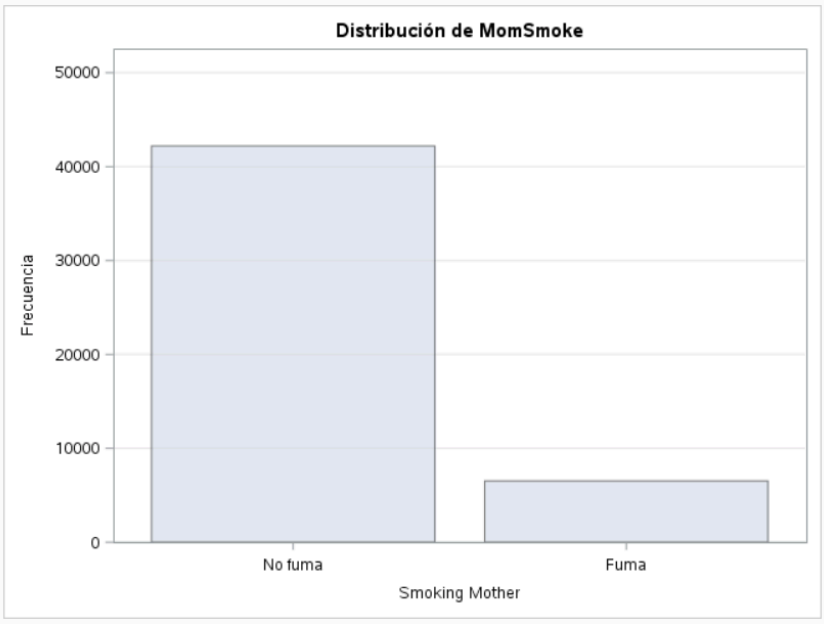
Estado Civil: Un 70% de madres casadas.

Married Mother		
Married	Frecuencia	Porcentaje
No Casada	14298	29.34
Casada	34436	70.66



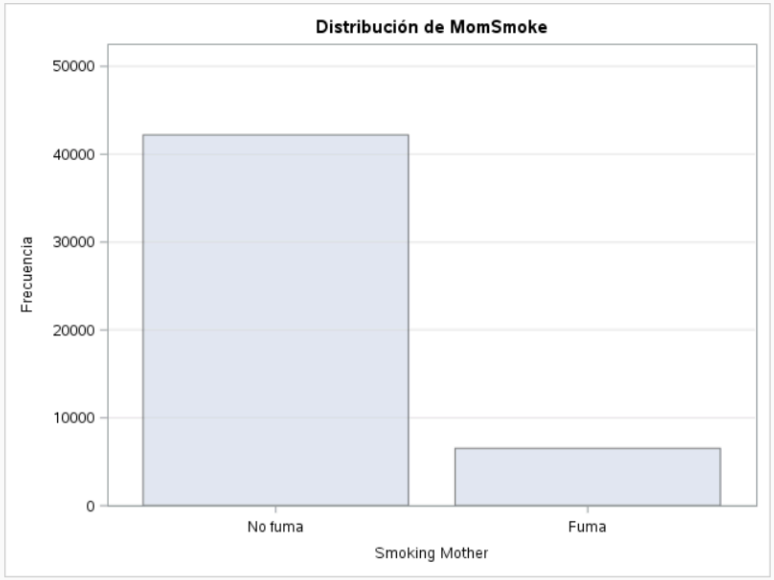
Fumadoras: Mayoría de madres no fumadoras (87%) frente a las que fuman (13%)

Smoking Mother		
MomSmoke	Frecuencia	Porcentaje
No fuma	42205	86.60
Fuma	6529	13.40



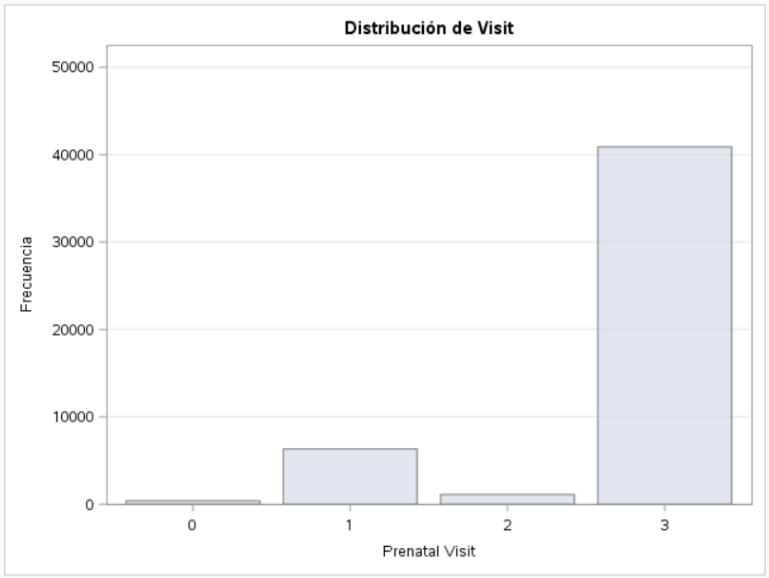
Nivel de estudios: en este caso, sí que me he atrevido a establecer las etiquetas a las 4 categorías. Hay una distribución variada de los cuatro casos posibles.

Smoking Mother		
MomSmoke	Frecuencia	Porcentaje
No fuma	42205	86.60
Fuma	6529	13.40



Visitas: hay claramente dos grupos o tres, los que han ido entre nunca y dos visitas, y las que han ido a las 3 visitas (84%). Podría agruparlas pero no sé el negocio y por tanto no voy a influir en el estudio agrupándolas sin conocimiento. O quizás sean trimestres (1,2, y 3) pero no lo sé.

Prenatal Visit		
Visit	Frecuencia	Porcentaje
0	403	0.83
1	6324	12.98
2	1113	2.28
3	40894	83.91



Variables contínuas, estádisticos básicos para ver sus observaciones, medias, mediana, mínimos y máximos y ver si hay missings

Estadisticos para las variables continuas (incluida objetivo)

Variable	Etiqueta	N	N Miss	Media	Mediana	Desv. est.	Mínimo	Máximo
Weight	Infant Birth Weight	48734	0	3368.21	3402.00	570.8653666	240.0000000	6350.00
xMomAge		48734	0	25.3823819	25.0000000	5.7561072	16.0000000	43.0000000
CigsPerDay	Cigarettes Per Day	48734	0	1.5141585	0	4.7073547	0	60.0000000
MomWtGain	Mother's Pregnancy Weight Gain	48734	0	0.6980547	0	12.9844704	-30.0000000	68.0000000

Vemos que hay un niño con 240 gramos de peso (quizás sea una anomalía del dataset), quizás lo veamos con más detalle más adelante.
Las edades de las madres parece que tienen sentido.
Los cigarrillos, pues entre las que no fuman y hay una que fuma 3 cajetillas al día.
El aumento de peso durante el embarazo, es bastante raro, si son kilos, hay una madre que perdió 30 kilos y otra que cogió 68 kilos, y la media se queda prácticamente en 0 kg de aumento de peso medio, parece un poco extraño.
Si se observa que no hay missings.

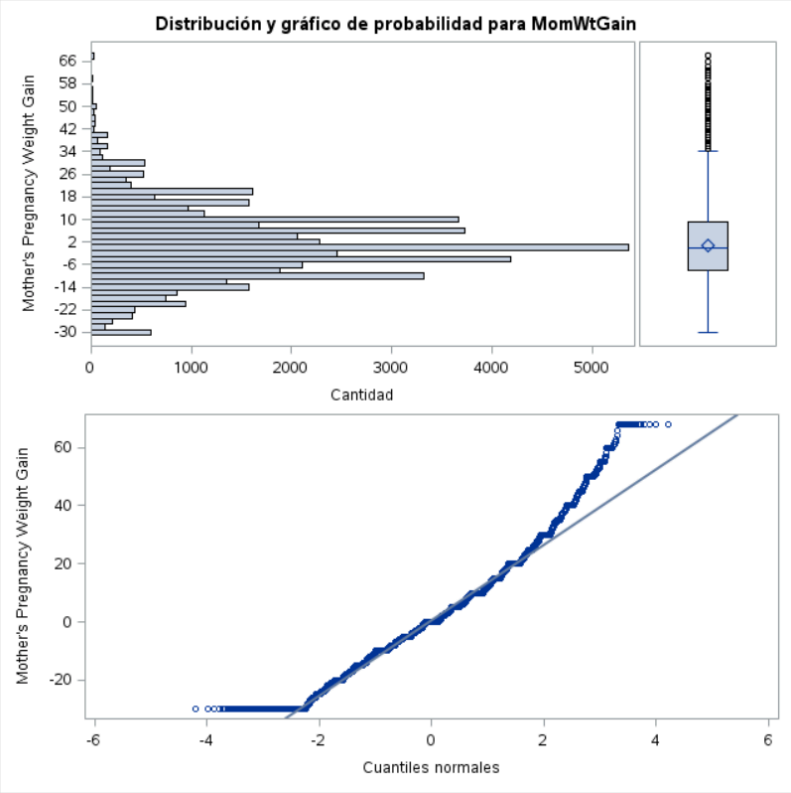
Como es rara la de aumento de peso, voy a estudiarla con más detalle.

Momentos

N	48734	Sumar pesos	48734
Media	0.69805475	Observ suma	34019
Desviación std	12.9844704	Varianza	168.596472
Asimetría	0.46271907	Curtosis	1.09275529
SC no corregida	8239959	SC corregida	8216211.88
Coef. variación	1860.09342	Media error std	0.05881772

Medidas estadísticas básicas

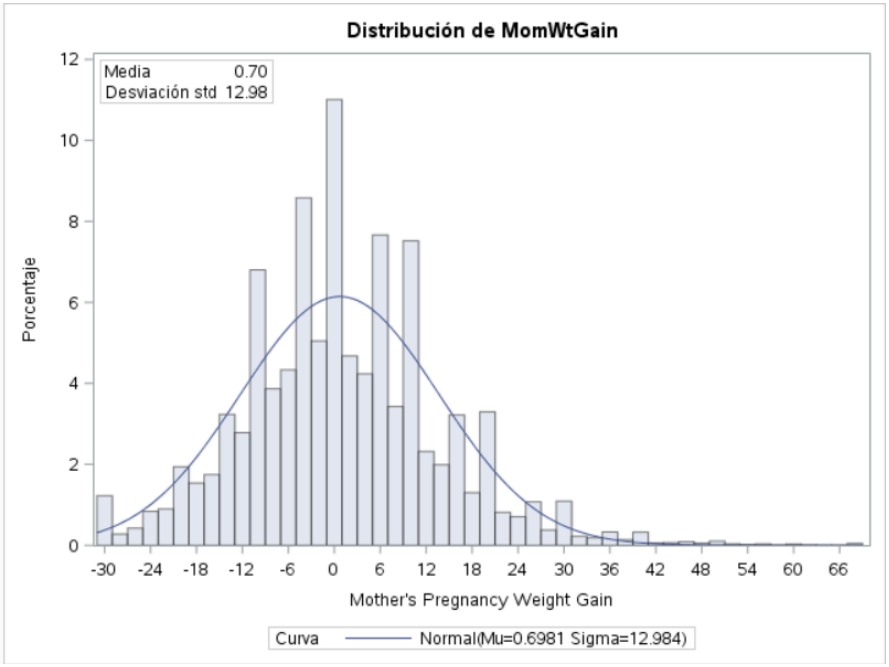
Ubicación		Variabilidad	
Media	0.698055	Desviación std	12.98447
Mediana	0.000000	Varianza	168.59647
Moda	0.000000	Rango	98.00000
		Rango intercuartil	17.00000



Se ve una cantidad apreciable de casos con sobrepeso por encima de la distribución

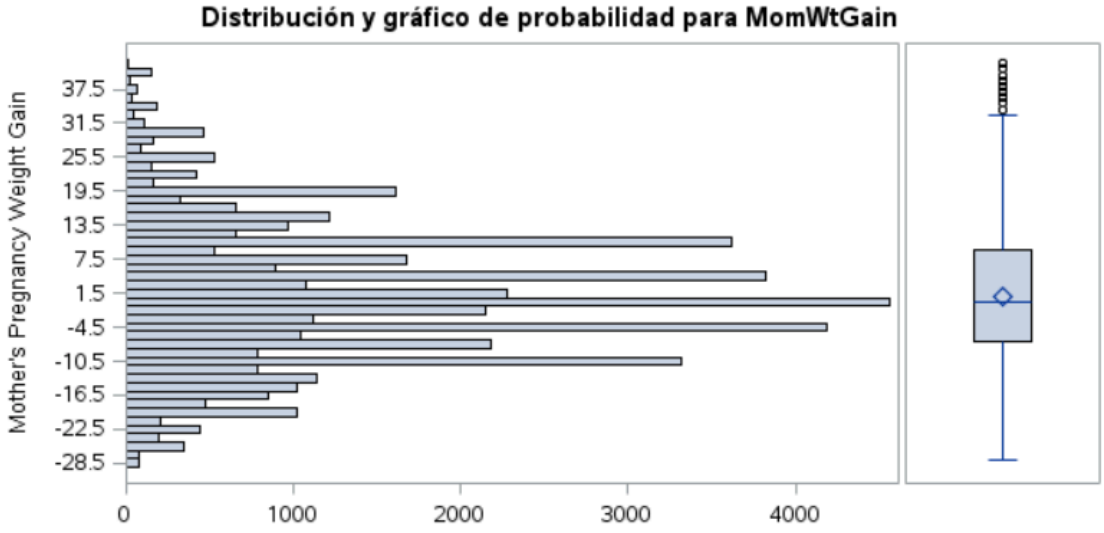
Cuantiles (Definición 5)	
Nivel	Cuantil
100% Máx	68
99%	37
95%	23
90%	17
75% Q3	9
50% Mediana	0
25% Q1	-8
10%	-15
5%	-20
1%	-30
0% Mín	-30

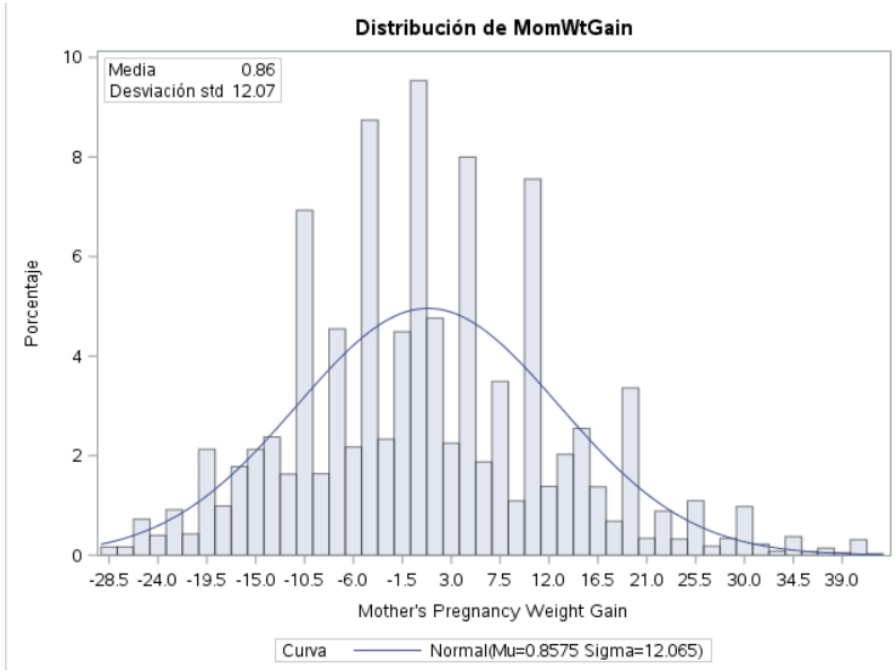
Observaciones extremas			
Inferior		Superior	
Valor	Obs	Valor	Obs
-30	48551	68	38648
-30	48485	68	39982
-30	48428	68	41291
-30	48126	68	41881
-30	47991	68	42071



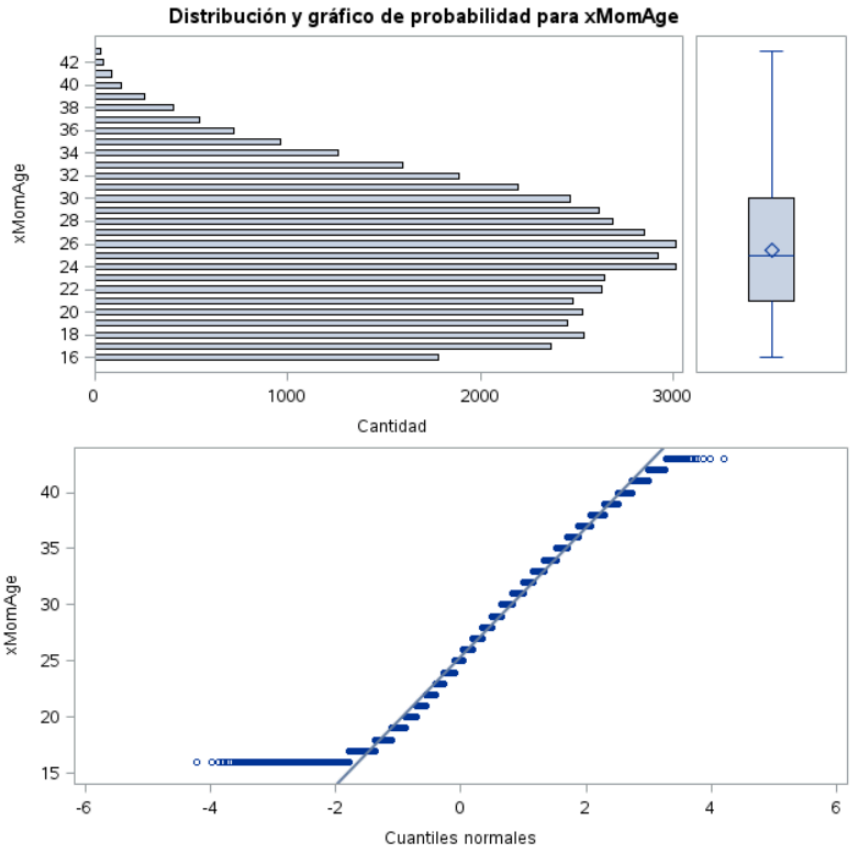
Es decir, por encima de 36 kg hay un 1% de registros, que deben rondar los 500 (al haber 50000 observaciones) y por el extremo inferior parece que hay otro 1% con pérdida de 29-30 kgs (otros 500 registros aproximadamente), creo que voy a optar por eliminarlos al considerarlos outliers, al menos por arriba encima de 42Kg y por debajo, los de -29 -30Kg.

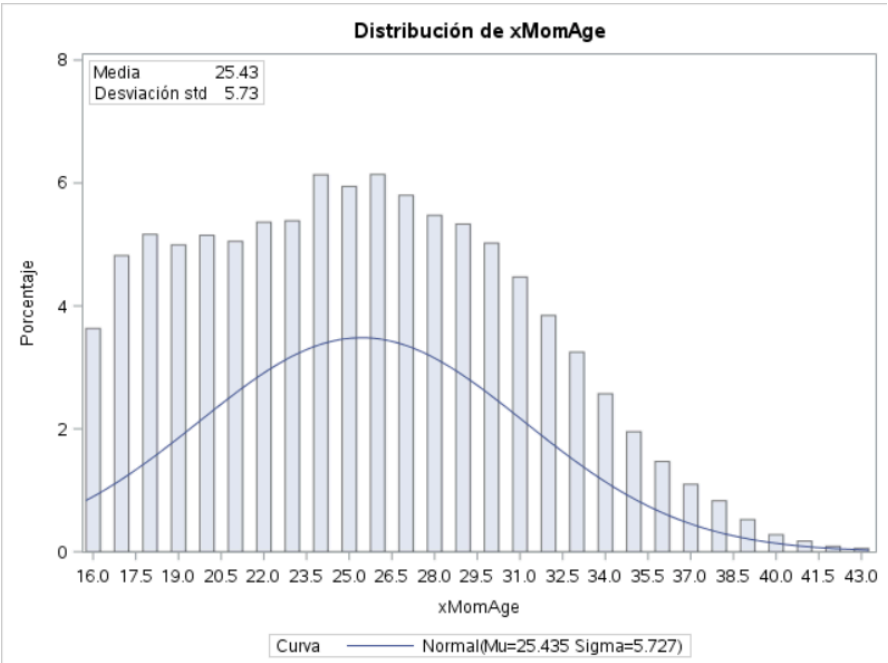
Veamos cómo queda ahora. Hemos eliminado unos 900 registros y la distribución es algo más normal. Aún así me quedan dudas de dónde cortar por debajo porque no sé si es normal tener pérdidas de peso tan notables.



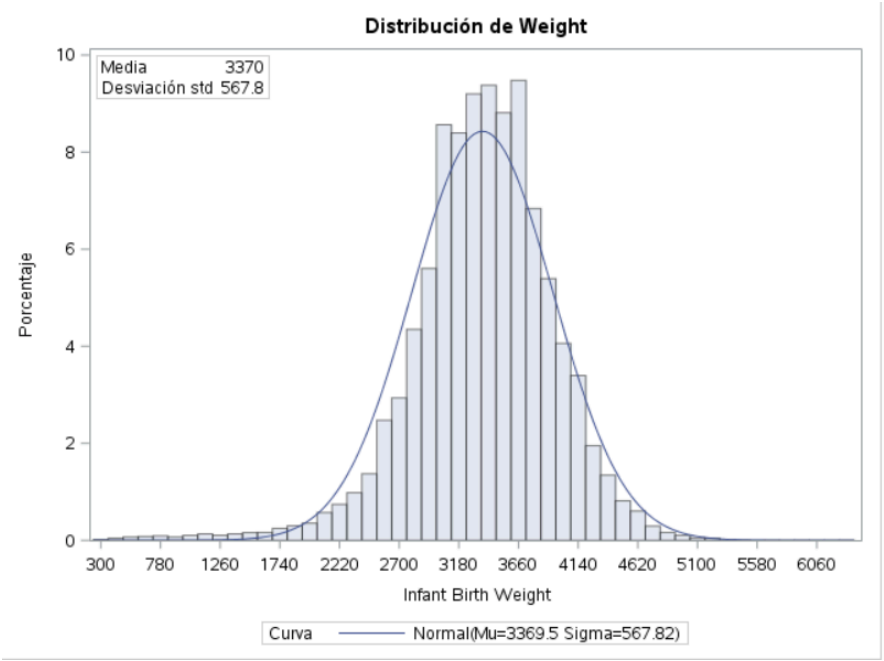

Distribución de MomWtGain
Media 0.86
Desviación std 12.07
Porcentaje
0
2
4
6
8
10
-28.5 -24.0 -19.5 -15.0 -10.5 -6.0 -1.5 3.0 7.5 12.0 16.5 21.0 25.5 30.0 34.5 39.0
Mother's Pregnancy Weight Gain
Curva Normal(Mu=0.8575 Sigma=12.065)

Mostramos ahora, los datos sobre la edad de las madres.





Ahora veamos cómo está de distribuida las variable objetivo. Weight.

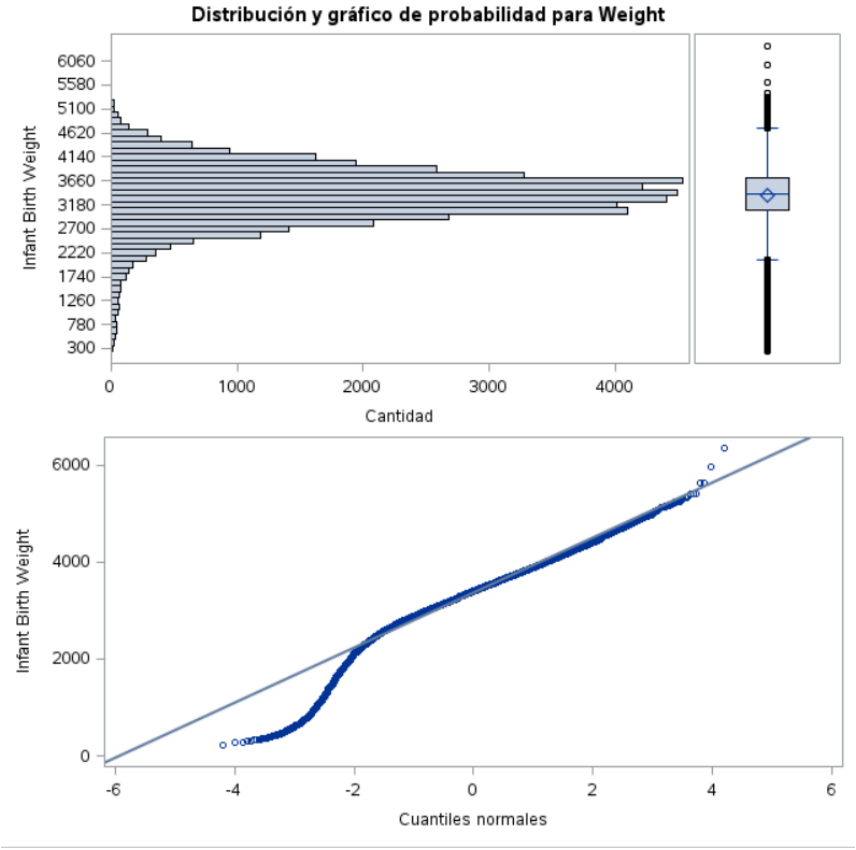


Variable: Weight (Infant Birth Weight)

Momentos			
N	47837	Sumar pesos	47837
Media	3369.53663	Observ suma	161188524
Desviación std	567.816543	Varianza	322415.627
Asimetría	-0.7302794	Curtosis	2.58049627
SC no corregida	5.58554E11	SC corregida	1.54231E10
Coef. variación	16.8514726	Media error std	2.59612783

Medidas estadísticas básicas			
Ubicación		Variabilidad	
Media	3369.537	Desviación std	567.81654
Mediana	3402.000	Varianza	322416
Moda	3402.000	Rango	6110
		Rango intercuartil	660.00000

Test para normalidad				
Test	Estadístico		P valor	
Kolmogorov-Smirnov	D	0.048551	Pr > D	<0.0100
Cramer-von Mises	W-Sq	30.01496	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	208.2233	Pr > A-Sq	<0.0050



Se ve claramente que hay un número significativo de niños con poco peso al nacer, por debajo de 1,9 Kg aproximadamente

Sin embargo esos pesos tan bajos de alrededor de 300 gramos, no creo que sean viables, por lo que consultando este artículo:
http://www.medigraphic.com/pdfs/inper/ip-2013/ip132b.pdf

Se ve que aquellos que nacieron antes de las 23 semanas de gestación tuvieron una media de 470 gramos y ninguno sobrevivió, y hasta las 28 semanas con un máximo de una media de 775 gramos, no tenían un esperanza de vida clara, pero dado que se indica que son nacidos, pondré el límite en 470 gramos, y por debajo de ese peso los elimino del conjunto de datos. Son 29 registros eliminados con ese peso menor de 470 gramos. Quedando los cuartiles como se muestra a continuación:

Cuantiles (Definición 5)	
Nivel	Cuantil
100% Máx	6350
99%	4607
95%	4224
90%	4026
75% Q3	3725
50% Mediana	3402
25% Q1	3062
10%	2722
5%	2466
1%	1616
0% Mín	475

Observaciones extremas			
Inferior		Superior	
Valor	Obs	Valor	Obs
475	1	5415	47804
482	6	5642	47805
482	5	5642	47806
482	4	5970	47807
482	3	6350	47808

Análisis de correlación.

Coeficientes de correlación Pearson, N = 47808
Prob > |r| suponiendo H0: Rho=0

	Weight	Black	Married	Boy	MomSmoke	CigsPerDay	MomWtGain	Visit	MomEdLevel	xMomAge
Weight Infant Birth Weight	1.00000	-0.15911 <.0001	0.15260 <.0001	0.10245 <.0001	-0.14425 <.0001	-0.12768 <.0001	0.20260 <.0001	0.06492 <.0001	-0.00039 0.9323	0.09916 <.0001
Black Black Mother	-0.15911 <.0001	1.00000	-0.35231 <.0001	-0.00403 0.3787	-0.04690 <.0001	-0.06435 <.0001	-0.05120 <.0001	-0.11648 <.0001	-0.05117 <.0001	-0.11568 <.0001
Married Married Mother	0.15260 <.0001	-0.35231 <.0001	1.00000	0.00720 0.1153	-0.17516 <.0001	-0.13379 <.0001	-0.00209 0.6470	0.21207 <.0001	0.01632 0.0004	0.36447 <.0001
Boy Baby Boy	0.10245 <.0001	-0.00403 0.3787	0.00720 0.1153	1.00000	0.00132 0.7725	0.00425 0.3531	0.02904 <.0001	0.00108 0.8132	-0.00357 0.4351	0.00018 0.9681
MomSmoke Smoking Mother	-0.14425 <.0001	-0.04690 <.0001	-0.17516 <.0001	0.00132 0.7725	1.00000	0.81795 <.0001	-0.01689 0.0002	-0.08919 <.0001	0.00699 0.1264	-0.08444 <.0001
CigsPerDay Cigarettes Per Day	-0.12768 <.0001	-0.06435 <.0001	-0.13379 <.0001	0.00425 0.3531	0.81795 <.0001	1.00000	-0.02826 <.0001	-0.07436 <.0001	0.01704 0.0002	-0.05319 <.0001
MomWtGain Mother's Pregnancy Weight Gain	0.20260 <.0001	-0.05120 <.0001	-0.00209 0.6470	0.02904 <.0001	-0.01689 0.0002	-0.02826 <.0001	1.00000	0.05095 <.0001	-0.03059 <.0001	-0.05759 <.0001
Visit Prenatal Visit	0.06492 <.0001	-0.11648 <.0001	0.21207 <.0001	0.00108 0.8132	-0.08919 <.0001	-0.07436 <.0001	0.05095 <.0001	1.00000	-0.04135 <.0001	0.14216 <.0001
MomEdLevel Mother's Education Level	-0.00039 0.9323	-0.05117 <.0001	0.01632 0.0004	-0.00357 0.4351	0.00699 0.1264	0.01704 0.0002	-0.03059 <.0001	-0.04135 <.0001	1.00000	0.04463 <.0001
xMomAge	0.09916 <.0001	-0.11568 <.0001	0.36447 <.0001	0.00018 0.9681	-0.08444 <.0001	-0.05319 <.0001	-0.05759 <.0001	0.14216 <.0001	0.04463 <.0001	1.00000

De este análisis sale claramente la correlación entre MomSmoke y CigsPerDay con un 82%, por lo que da lo mismo los cigarrillos que fume la madre, parece que lo que importa es lo que fuma, dado que Weight con MomSmoke y con CigsPerDay da valores similares 0.14, 0.12.

2. MODELO

Con el modelo básico:
Weight = Married Boy Black xMomAge Visit MomSmoke MomEdLevel MomWtGain

R-cuadrado	Var Coef.	Raíz MSE	Media de Weight
0.109183	15.78469	532.1571	3371.349

Origen	DF	Tipo III SS	Cuadrado de la media	Valor F	Pr > F
Married	1	23962865.9	23962865.9	84.62	<.0001
Boy	1	139325000.4	139325000.4	491.98	<.0001
Black	1	203349998.0	203349998.0	718.07	<.0001
xMomAge	27	53380209.7	1977044.8	6.98	<.0001
Visit	3	6554594.7	2184864.9	7.72	<.0001
MomSmoke	1	219643622.5	219643622.5	775.60	<.0001
MomEdLevel	3	7382879.2	2460959.7	8.69	<.0001
MomWtGain	70	601562095.6	8593744.2	30.35	<.0001

Buscamos el mejor modelo ejecutando en la macro todas las combinaciones posibles entre variables

El resultado final, es que destacan los cuatro de la lista sobre el resto

	modelo	COUNT	PERCENT
1	Intercept Black Married Boy MomSmoke MomWtGain MomEdLevel	209	29.027777778
2	Intercept Black Married Boy MomWtGain Visit MomEdLevel MomSmoke*xMomAge	153	21.25
3	Intercept Married Boy MomWtGain Visit MomEdLevel Black*xMomAge MomSmoke*xMomAge	132	18.333333333
4	Intercept Black Boy MomWtGain Visit MomEdLevel Married*xMomAge MomSmoke*xMomAge	103	14.305555556
5	Intercept Boy MomWtGain Visit MomEdLevel Black*xMomAge Married*xMomAge MomSmoke*xMomAge	26	3.6111111111
6	Intercept Black Married Boy MomWtGain MomSmoke*xMomAge	16	2.2222222222
7	Intercept Black Boy MomWtGain Married*xMomAge MomSmoke*xMomAge	15	2.0833333333
8	Intercept Married Boy MomWtGain Black*xMomAge MomSmoke*xMomAge	11	1.5277777778
9	Intercept Black Married Boy MomWtGain MomEdLevel MomSmoke*xMomAge	9	1.25
10	Intercept Married Boy MomSmoke MomWtGain Visit MomEdLevel Black*xMomAge	8	1.1111111111
11	Intercept Black Boy MomWtGain Visit Married*xMomAge MomSmoke*xMomAge	7	0.9722222222
12	Intercept Married Boy MomWtGain MomEdLevel Black*xMomAge MomSmoke*xMomAge	7	0.9722222222
13	Intercept Black Boy MomWtGain MomEdLevel Married*xMomAge MomSmoke*xMomAge	6	0.8333333333
14	Intercept Boy MomSmoke MomWtGain Visit MomEdLevel Black*xMomAge Married*xMomAge	6	0.8333333333
15	Intercept Married Boy MomWtGain Visit Black*xMomAge MomSmoke*xMomAge	5	0.6944444444
16	Intercept Black Married Boy MomWtGain Visit MomSmoke*xMomAge	2	0.2777777778
17	Intercept Boy MomSmoke MomWtGain Black*xMomAge Married*xMomAge	2	0.2777777778
18	Intercept Boy MomWtGain Visit Black*xMomAge Married*xMomAge MomSmoke*xMomAge	2	0.2777777778
19	Intercept Married Boy MomSmoke MomWtGain Visit Black*xMomAge	1	0.1388888889

- Black M Black Married Boy MomSmoke MomWtGain MomEdLevel
- Black Married Boy MomWtGain Visit MomEdLevel MomSmoke*xMomAge

- Married Boy MomWtGain Visit MomEdLevel Black*xMomAge MomSmoke*xMomAge

- Black Boy MomWtGain Visit MomEdLevel Married*xMomAge MomSmoke*xMomAge

Vamos a ver sus prestaciones

- Black M Black Married Boy MomSmoke MomWtGain MomEdLevel

R-cuadrado	Var Coef.	Raíz MSE	Media de Weight
0.105289	15.81418	533.1513	3371.349

Origen	DF	Tipo III SS	Cuadrado de la media	Valor F	Pr > F
Black	1	202447988.6	202447988.6	712.22	<.0001
Married	1	54627743.7	54627743.7	192.18	<.0001
Boy	1	139755825.0	139755825.0	491.66	<.0001
MomSmoke	1	212758795.9	212758795.9	748.49	<.0001
MomWtGain	70	582558511.6	8322264.5	29.28	<.0001
MomEdLevel	3	29177003.7	9725667.9	34.22	<.0001

Black Married Boy MomWtGain Visit MomEdLevel MomSmoke*xMomAge

R-cuadrado	Var Coef.	Raíz MSE	Media de Weight
0.111125	15.77194	531.7272	3371.349

Origen	DF	Tipo III SS	Cuadrado de la media	Valor F	Pr > F
Black	1	196679097.7	196679097.7	695.63	<.0001
Married	1	23292606.6	23292606.6	82.38	<.0001
Boy	1	138488572.0	138488572.0	489.82	<.0001
MomWtGain	70	600958972.8	8585128.2	30.36	<.0001
Visit	3	5998803.2	1999601.1	7.07	<.0001
MomEdLevel	3	6720568.7	2240189.6	7.92	<.0001
MomSmoke*xMomAge	55	293280625.4	5332375.0	18.86	<.0001

Married Boy MomWtGain Visit MomEdLevel Black*xMomAge MomSmoke*xMomAge

R-cuadrado	Var Coef.	Raíz MSE	Media de Weight
0.111968	15.76892	531.6254	3371.349

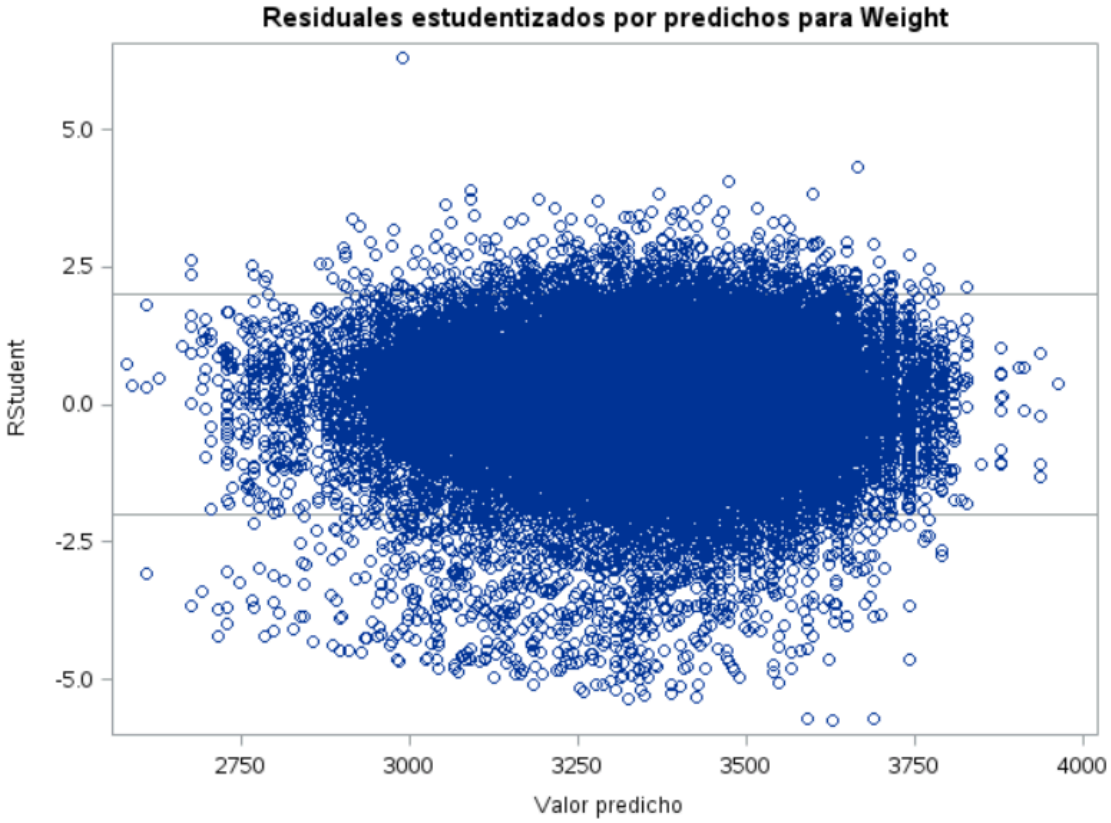
Origen	DF	Tipo III SS	Cuadrado de la media	Valor F	Pr > F
Married	1	23989835.4	23989835.4	84.88	<.0001
Boy	1	138106488.8	138106488.8	488.66	<.0001
MomWtGain	70	603188440.6	8616977.7	30.49	<.0001
Visit	3	6229483.2	2076494.4	7.35	<.0001
MomEdLevel	3	6369751.6	2123250.5	7.51	<.0001
Black*xMomAge	28	209468337.6	7481012.1	26.47	<.0001
MomSmoke*xMomAge	28	243204303.7	8685868.0	30.73	<.0001

Black Boy MomWtGain Visit MomEdLevel Married*xMomAge MomSmoke*xMomAge

R-cuadrado	Var Coef.	Raíz MSE	Media de Weight
0.112316	15.76584	531.5215	3371.349

Origen	DF	Tipo III SS	Cuadrado de la media	Valor F	Pr > F
Black	1	196381930.0	196381930.0	695.12	<.0001
Boy	1	138094112.9	138094112.9	488.80	<.0001
MomWtGain	70	602618143.3	8608830.6	30.47	<.0001
Visit	3	5977862.2	1992620.7	7.05	<.0001
MomEdLevel	3	5758051.0	1919350.3	6.79	0.0001
Married*xMomAge	28	41348750.5	1476741.1	5.23	<.0001
MomSmoke*xMomAge	28	236948129.3	8462433.2	29.95	<.0001

Por lo tanto parece que el primer modelo, el que tiene un R2 de 10,53% es el mejor que he conseguido, no es demasiado bueno. Mirando la gráfica se ve el ajuste que hace en la predicción.



SEGUNDA PARTE. CLASIFICACIÓN

Objetivo: identificar a la población que es más propensa a sufrir el efecto de la mediación del bajo peso al nacer sobre la mortalidad infantil.
l

Datos:

LowBirthWgt	Low Birth Weigt	Bajo peso al nacer (Target dicotómica: yes, no)

AgeGroup	Mother's Age Group	Grupo de edad (niveles:1, 2, 3)
Death	Death	Muerte (yes, no)
Drinking	Mother's Drinking	Madre alcoholica (yes, no)
Married	Married Mother	Madre casada (yes,no)
Race	Race	Raza (niveles: asian, black, hispanic, native, white)
Smoking	Smoking Mother	Madre Fumadora (yes, no)
SomeCollege	SomeCollege	Alguna educación superior (yes, no)

1. Análisis exploratorio de los datos

Tras la carga de los datos, 100.000 observaciones, mostramos los primeros registros de la tabla.

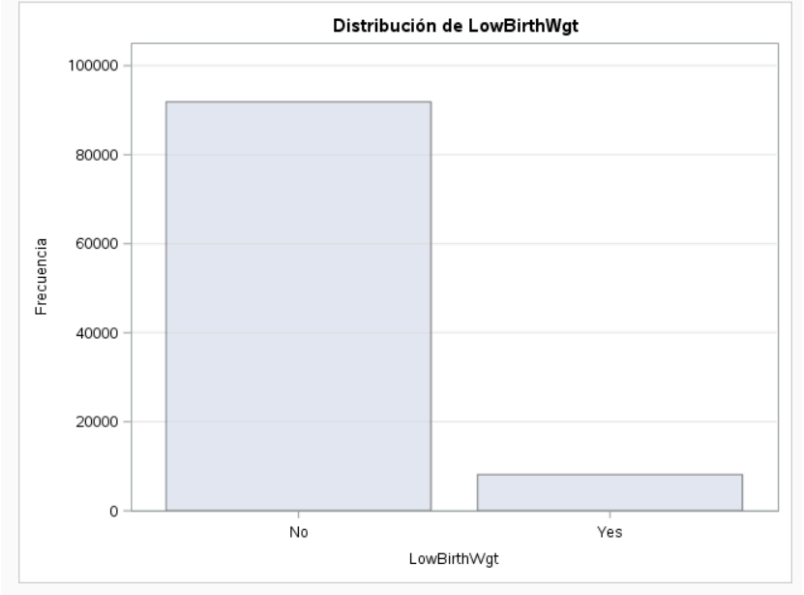
Primeros registros del dataset

Obs	LowBirthWgt	Married	AgeGroup	Race	Drinking	Death	Smoking	SomeCollege
1	No	No	3	Asian	No	No	No	Yes
2	No	No	2	White	No	No	No	No
3	Yes	Yes	2	Native	No	Yes	No	No
4	No	No	2	White	No	No	No	No
5	No	No	2	White	No	No	No	Yes
6	No	No	2	White	No	No	No	
7	No	No	2	Asian	No	No	No	Yes
8	No	No	3	White	No	No	No	Yes
9	No	Yes	1	Black	No	No	No	No
10	No	No	2	Native	No	No	No	Yes
11	No	No	2	White	No	No	No	Yes
12	No	No	2	White	No	No	No	Yes
13	No	No	2	Native	No	No	No	No
14	No	No	2	White	No	No	No	No
15	No	No	3	White	No	No	No	Yes

Todas las variables son categóricas, y ya ha asoma algún missing en esta primera vista. Vamos a verlo con detalle para cada una de las variables.

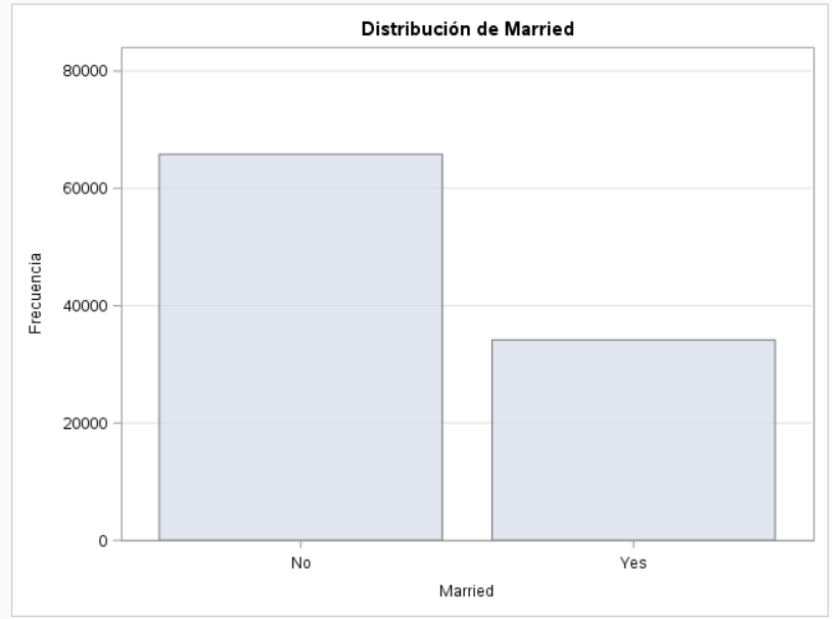
Bajo peso al nacer: en el dataset predominan los niños con peso normal.

LowBirthWgt	Frecuencia	Porcentaje
No	91859	91.86
Yes	8141	8.14



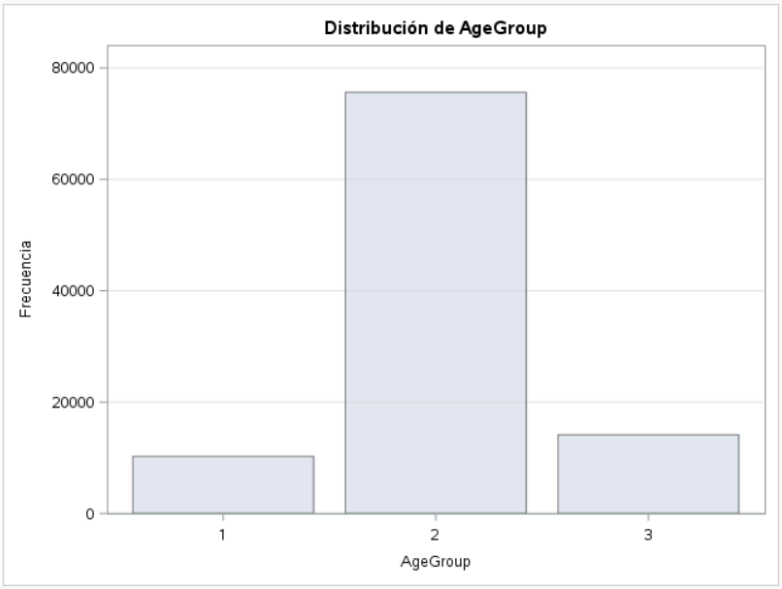
Estado civil: predominan las madre no casadas.

Married	Frecuencia	Porcentaje
No	65830	65.83
Yes	34170	34.17



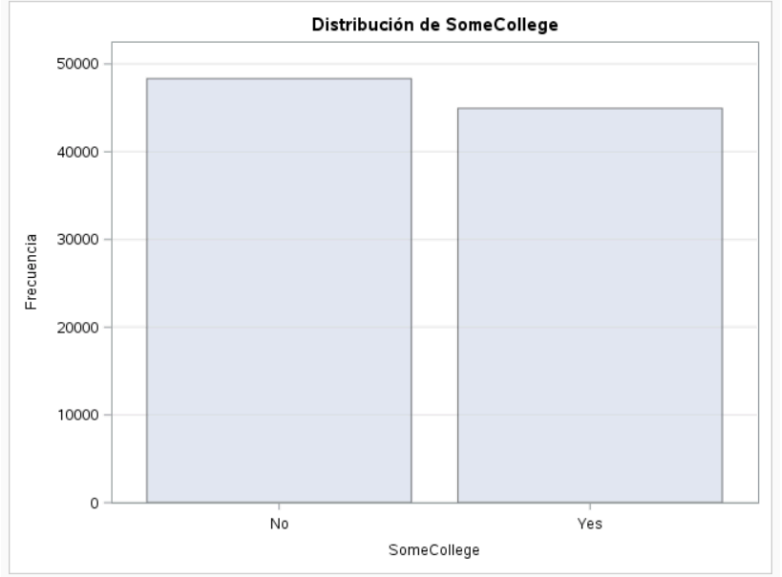
Grupo de edad: predomina claramente los del grupo 2 de edad, supongo que serían los que están entorno a 25-30 años.

AgeGroup	Frecuencia	Porcentaje
1	10245	10.25
2	75633	75.63
3	14122	14.12



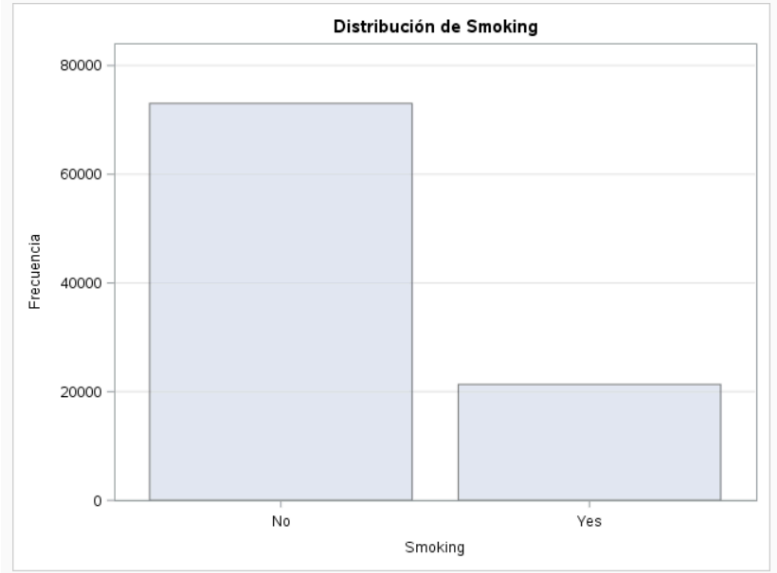
Nivel de educación: si han pasado por estudios superiores o no. En este caso, hay 6708 missings, proporcionalmente, sobre 100.000 registros es una proporción escasa, los voy a repartir al 50% entre ambas clases, por lo que la distribución quedará similar a la actual.

SomeCollege	Frecuencia	Porcentaje
No	48339	51.81
Yes	44953	48.19
Frequency Missing = 6708		



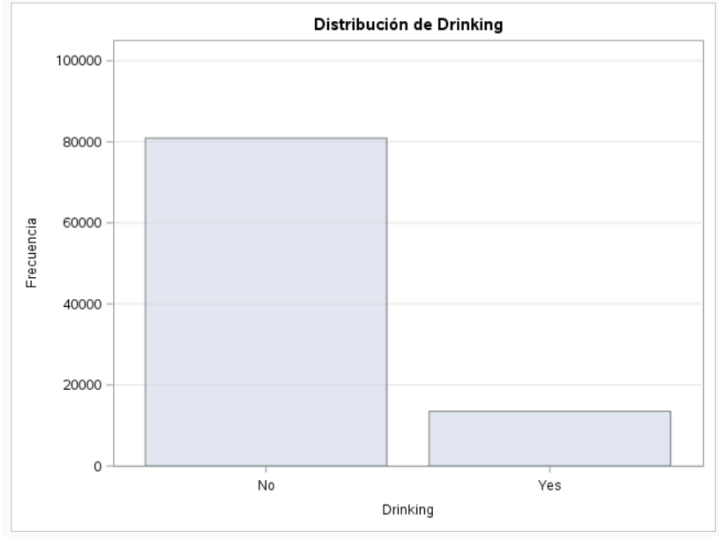
Fumadoras: en este caso claramente predominan los no fumadores, en este caso los missing, los voy a asignar al mayoritario, al No.

Smoking	Frecuencia	Porcentaje
No	73043	77.39
Yes	21345	22.61
Frequency Missing = 5612		



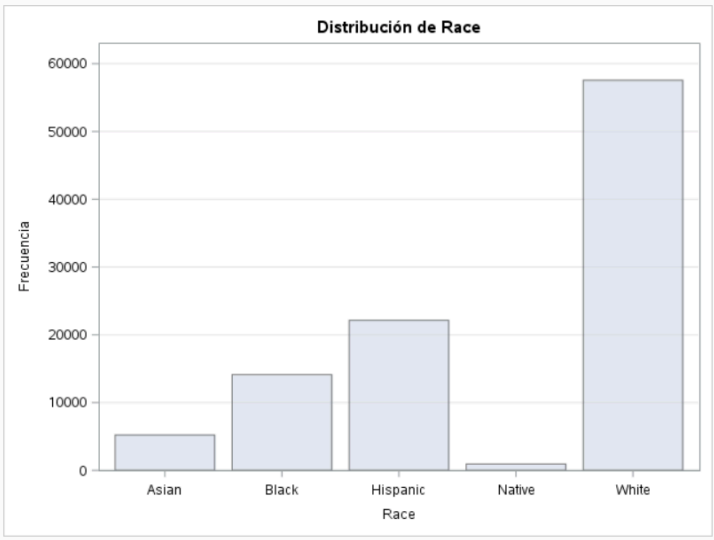
Bebida: en este caso y con esta distribución de frecuencias, los missings se los voy a asignar al No.

Drinking	Frecuencia	Porcentaje
No	80914	85.72
Yes	13474	14.28
Frequency Missing = 5612		



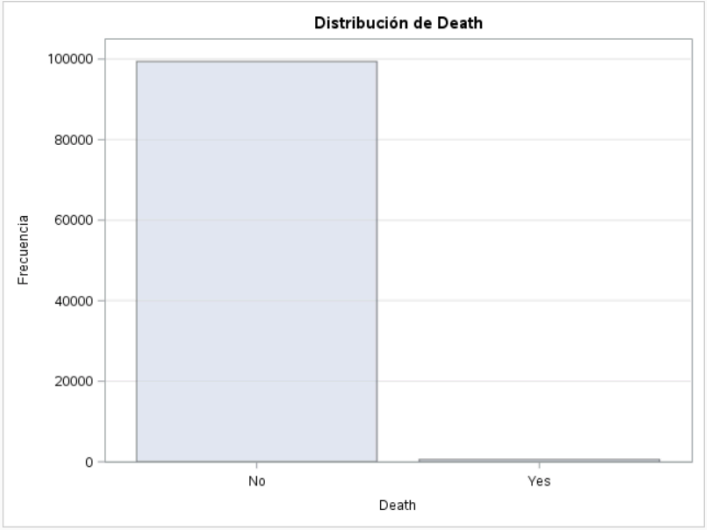
Raza:no tenemos missings, y hay una distribución bastante dispar con una prevalencia de raza blanca y prácticamente irrelevante en la Nativa Americana.

Race	Frecuencia	Porcentaje
Asian	5224	5.22
Black	14133	14.13
Hispanic	22139	22.14
Native	942	0.94
White	57562	57.56



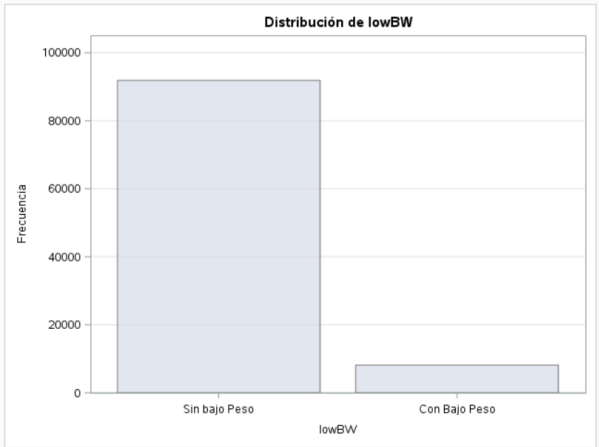
Muerte: en este dataset el apenas el 0,6% de los niños mueren tras nacer.

Death	Frecuencia	Porcentaje
No	99411	99.41
Yes	589	0.59

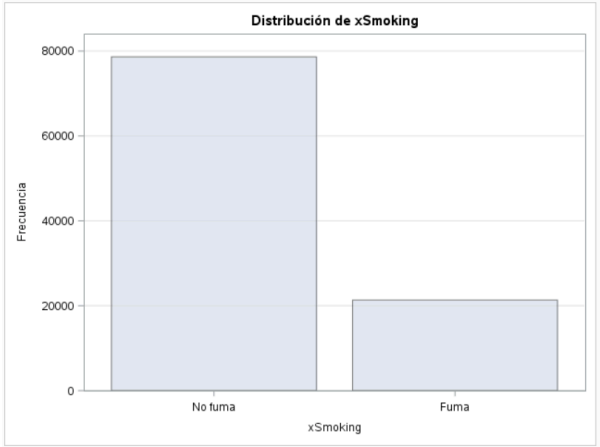


Proceso de *dicotomización* de las variables, en este paso, *dicotomizo* las diferentes clases de todas las variables categóricas y hago el reparto de los missings que he explicado en el punto anterior. También he etiquetado, mediante formato, las clases a la hora de mostrar sus frecuencias para que sean más entendibles.

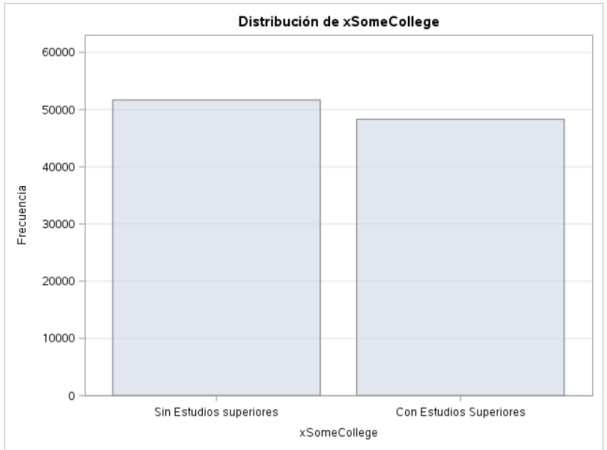
lowBW	Frecuencia	Porcentaje
Sin bajo Peso	91859	91.86
Con Bajo Peso	8141	8.14



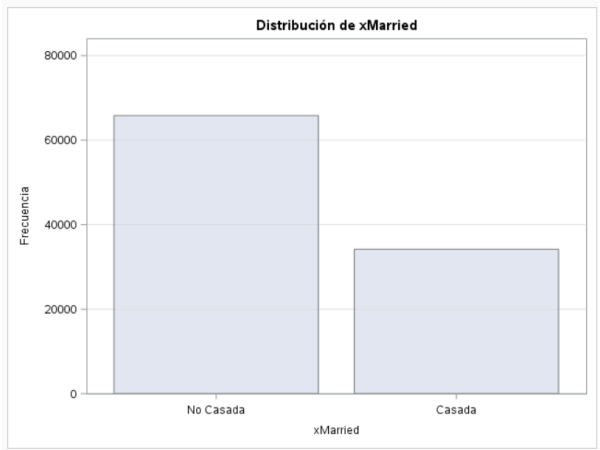
xSmoking	Frecuencia	Porcentaje
No fuma	78655	78.66
Fuma	21345	21.35



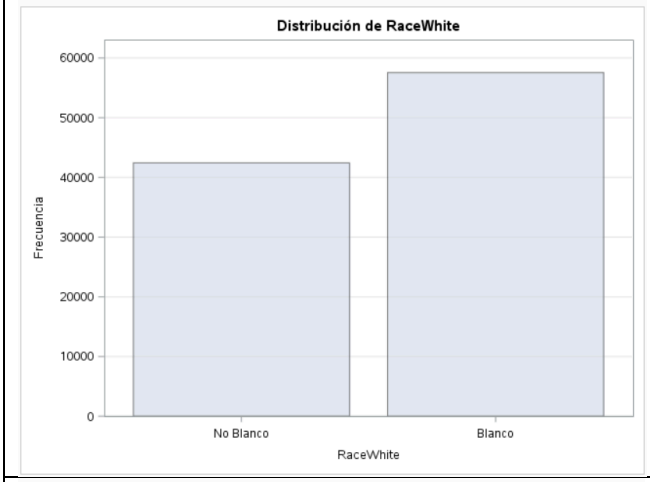
xSomeCollege	Frecuencia	Porcentaje
Sin Estudios superiores	51676	51.68
Con Estudios Superiores	48324	48.32



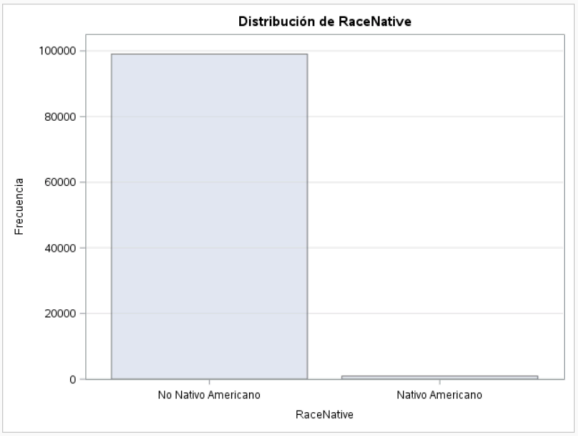
xMarried	Frecuencia	Porcentaje
No Casada	65830	65.83
Casada	34170	34.17



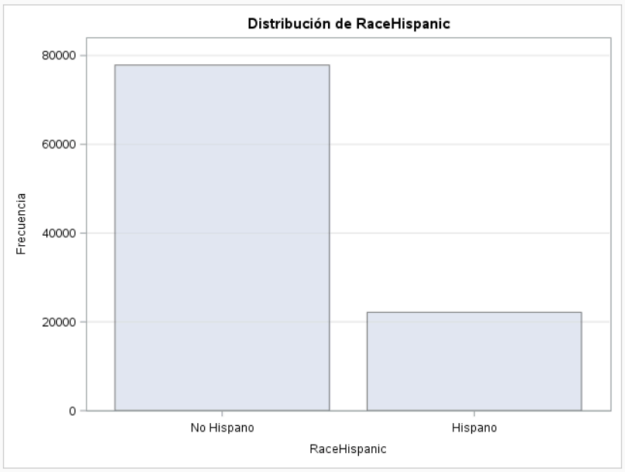
RaceWhite	Frecuencia	Porcentaje
No Blanco	42438	42.44
Blanco	57562	57.56



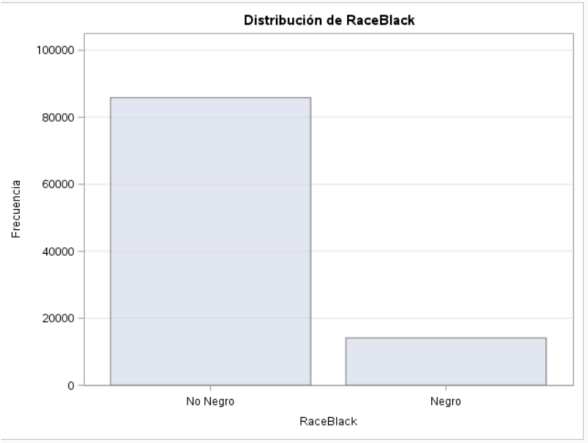
RaceNative	Frecuencia	Porcentaje
No Nativo Americano	99058	99.06
Nativo Americano	942	0.94



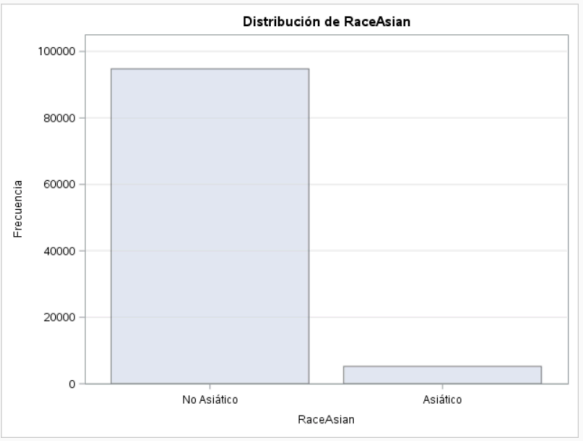
RaceHispanic	Frecuencia	Porcentaje
No Hispano	77861	77.86
Hispano	22139	22.14



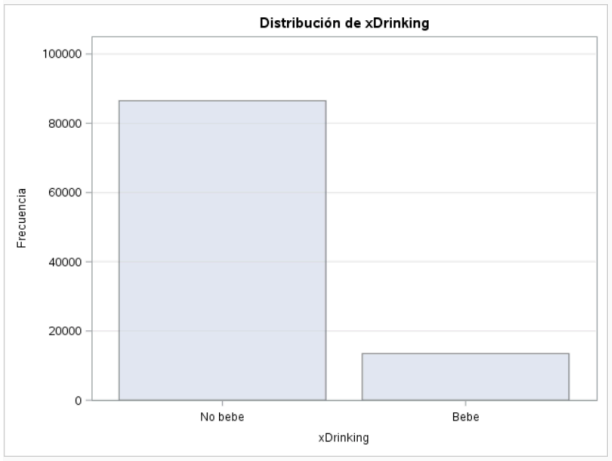
RaceBlack	Frecuencia	Porcentaje
No Negro	85867	85.87
Negro	14133	14.13



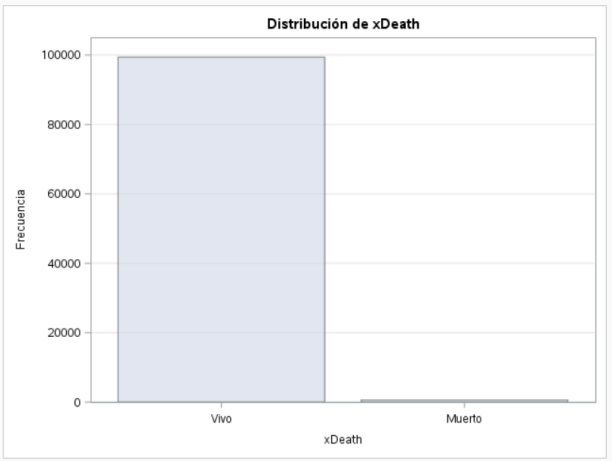
RaceAsian	Frecuencia	Porcentaje
No Asiático	94776	94.78
Asiático	5224	5.22



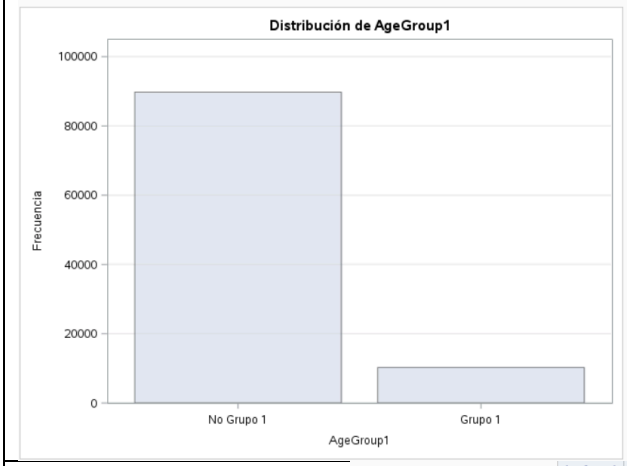
xDrinking	Frecuencia	Porcentaje
No bebe	86526	86.53
Bebe	13474	13.47



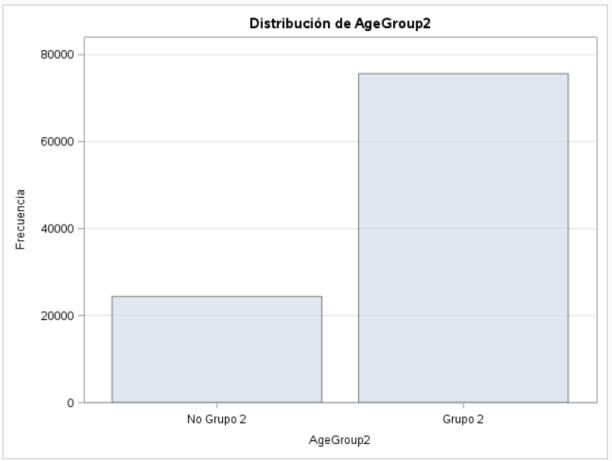
xDeath	Frecuencia	Porcentaje
Vivo	99411	99.41
Muerto	589	0.59



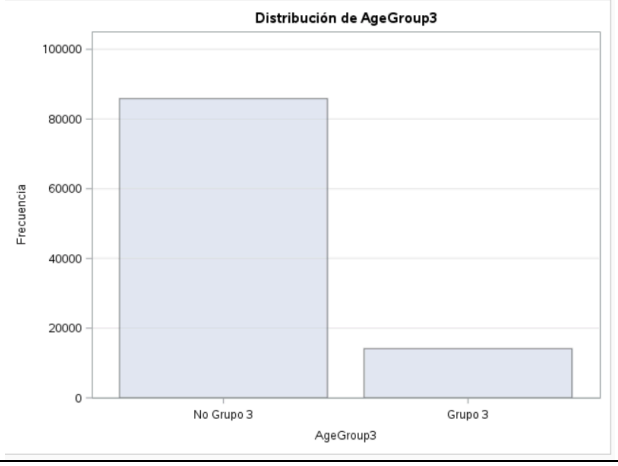
AgeGroup1	Frecuencia	Porcentaje
No Grupo 1	89755	89.76
Grupo 1	10245	10.25



AgeGroup2	Frecuencia	Porcentaje
No Grupo 2	24367	24.37
Grupo 2	75633	75.63



AgeGroup3	Frecuencia	Porcentaje
No Grupo 3	85878	85.88
Grupo 3	14122	14.12



Ordenamos el dataset y eliminamos los duplicados, y nos quedamos apenas con unas 500 observaciones.

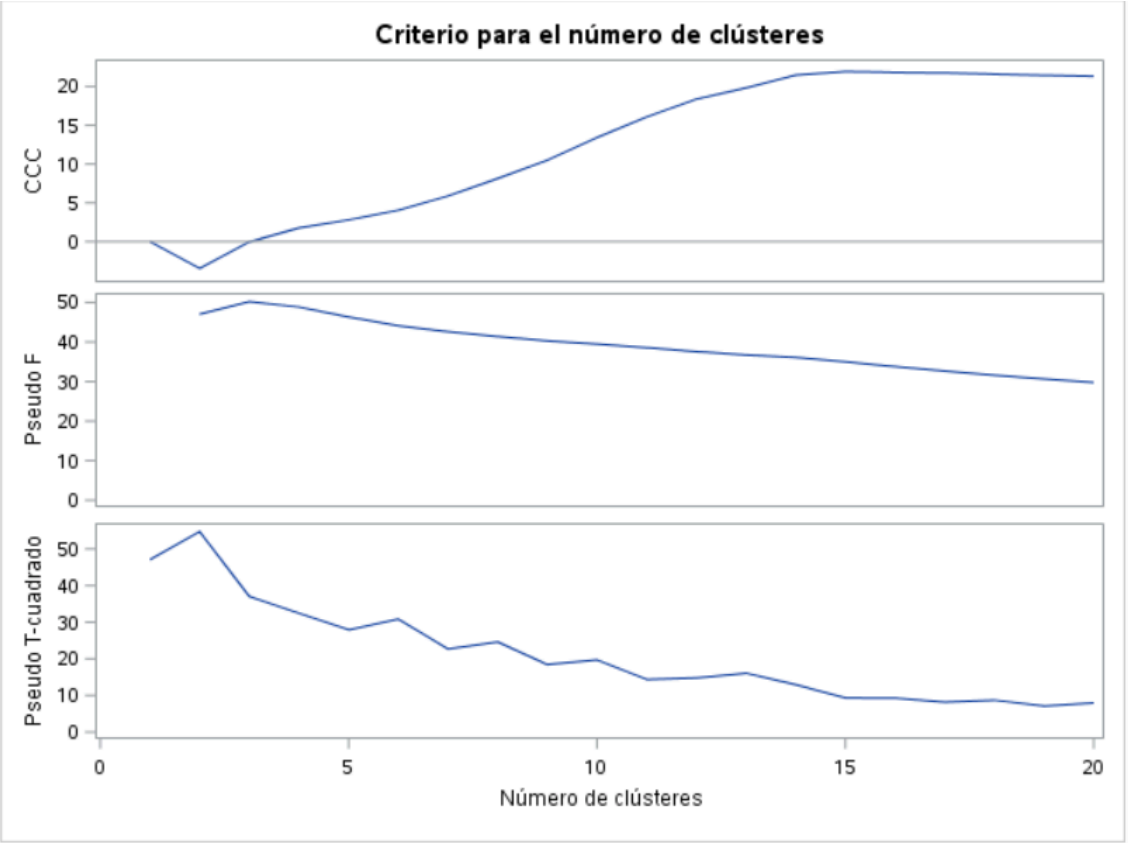
Análisis de Correlación. No hay correlaciones relevantes, una leve entre Smoking y Drinking.

Coeficientes de correlación Pearson, N = 483
Prob > |r| suponiendo H0: Rho=0

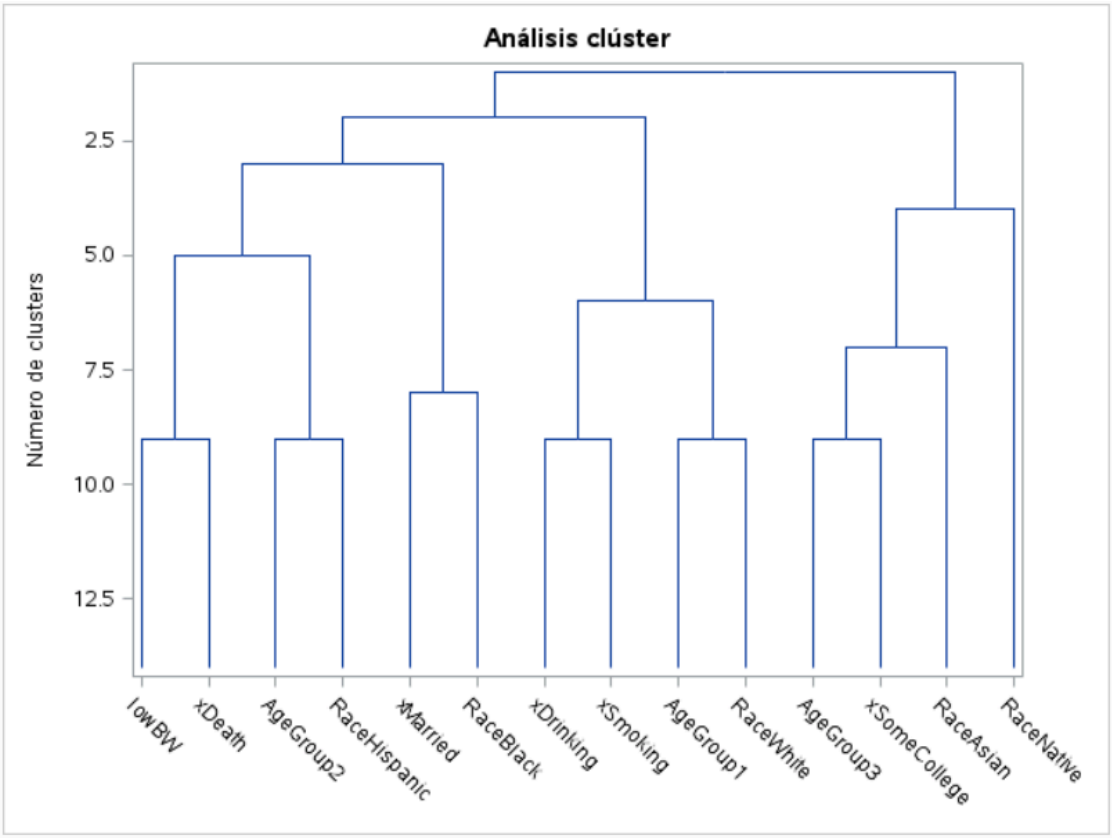
	lowBW	AgeGroup1	AgeGroup2	AgeGroup3	xDeath	xMarried	RaceAsian	RaceBlack	RaceHispanic	RaceNative	RaceWhite	xDrinking	xSomeCollege	xSmoking
lowBW	1.00000	-0.05856 0.1989	0.00986 0.8288	0.04363 0.3386	0.16062 0.0004	0.05555 0.2230	-0.00394 0.9312	0.02448 0.5915	-0.00977 0.8304	-0.07410 0.1038	0.04396 0.3351	-0.07501 0.0996	-0.00898 0.8439	0.02480 0.5867
AgeGroup1	-0.05856 0.1989	1.00000	-0.49227 <.0001	-0.40501 <.0001	-0.07398 0.1044	0.10167 0.0255	0.00927 0.8390	-0.01511 0.7405	-0.02358 0.6052	-0.02598 0.5690	0.04728 0.2997	-0.04513 0.3223	-0.11660 0.0103	-0.00602 0.8950
AgeGroup2	0.00986 0.8288	-0.49227 <.0001	1.00000	-0.59648 <.0001	0.09169 0.0440	-0.01693 0.7105	-0.02707 0.5529	0.00884 0.8463	0.01619 0.7226	0.04103 0.3683	-0.03109 0.4954	0.04160 0.3616	0.04986 0.2742	-0.01740 0.7029
AgeGroup3	0.04363 0.3386	-0.40501 <.0001	-0.59648 <.0001	1.00000	-0.02810 0.5379	-0.07596 0.0954	0.01988 0.6629	0.00464 0.9189	0.00473 0.9174	-0.01914 0.6748	-0.01094 0.8105	-0.00208 0.9636	0.05514 0.2264	0.02382 0.6015
xDeath	0.16062 0.0004	-0.07398 0.1044	0.09169 0.0440	-0.02810 0.5379	1.00000	0.01343 0.7685	-0.06740 0.1391	0.00446 0.9221	0.04066 0.3726	-0.15099 0.0009	0.12359 0.0065	-0.12474 0.0060	-0.08286 0.0688	-0.08221 0.0710
xMarried	0.05555 0.2230	0.10167 0.0255	-0.01693 0.7105	-0.07596 0.0954	0.01343 0.7685	1.00000	-0.02234 0.6243	0.01137 0.8031	-0.00464 0.9190	0.03303 0.4689	-0.01205 0.7917	0.00026 0.9954	-0.02344 0.6073	0.03303 0.4690
RaceAsian	-0.00394 0.9312	0.00927 0.8390	-0.02707 0.5529	0.01988 0.6629	-0.06740 0.1391	-0.02234 0.6243	1.00000	-0.22953 <.0001	-0.23371 <.0001	-0.16245 0.0003	-0.26985 <.0001	0.01567 0.7313	0.03650 0.4235	-0.03232 0.4786
RaceBlack	0.02448 0.5915	-0.01511 0.7405	0.00884 0.8463	0.00464 0.9189	0.00446 0.9221	0.01137 0.8031	-0.22953 <.0001	1.00000	-0.28284 <.0001	-0.19660 <.0001	-0.32658 <.0001	-0.02511 0.5819	-0.00088 0.9846	-0.00349 0.9390
RaceHispanic	-0.00977 0.8304	-0.02358 0.6052	0.01619 0.7226	0.00473 0.9174	0.04066 0.3726	-0.00464 0.9190	-0.23371 <.0001	-0.28284 <.0001	1.00000	-0.20019 <.0001	-0.33253 <.0001	0.06545 0.1509	-0.02407 0.5977	0.00099 0.9827
RaceNative	-0.07410 0.1038	-0.02598 0.5690	0.04103 0.3683	-0.01914 0.6748	-0.15099 0.0009	0.03303 0.4689	-0.16245 0.0003	-0.19660 <.0001	-0.20019 <.0001	1.00000	-0.23114 <.0001	-0.07433 0.1028	-0.04964 0.2762	-0.00681 0.8813
RaceWhite	0.04396 0.3351	0.04728 0.2997	-0.03109 0.4954	-0.01094 0.8105	0.12359 0.0065	-0.01205 0.7917	-0.26985 <.0001	-0.32658 <.0001	-0.33253 <.0001	-0.23114 <.0001	1.00000	0.00378 0.9340	0.02967 0.5153	0.03370 0.4599
xDrinking	-0.07501 0.0996	-0.04513 0.3223	0.04160 0.3616	-0.00208 0.9636	-0.12474 0.0060	0.00026 0.9954	0.01567 0.7313	-0.02511 0.5819	0.06545 0.1509	-0.07433 0.1028	0.00378 0.9340	1.00000	0.00309 0.9460	0.34677 <.0001
xSomeCollege	-0.00898 0.8439	-0.11660 0.0103	0.04986 0.2742	0.05514 0.2264	-0.08286 0.0688	-0.02344 0.6073	0.03650 0.4235	-0.00088 0.9846	-0.02407 0.5977	-0.04964 0.2762	0.02967 0.5153	0.00309 0.9460	1.00000	0.00225 0.9607
xSmoking	0.02480 0.5867	-0.00602 0.8950	-0.01740 0.7029	0.02382 0.6015	-0.08221 0.0710	0.03303 0.4690	-0.03232 0.4786	-0.00349 0.9390	0.00099 0.9827	-0.00681 0.8813	0.03370 0.4599	0.34677 <.0001	0.00225 0.9607	1.00000

Cluster mediante el método centroide, parece que en el entorno de 8-9 estaría una posible cantidad de clusters a considerar.

Cluster mediante el Ward.



Varclus, pongamos 9 como el número de clusters que parece que sería el apropiado, tenemos el siguiente resultado.


Análisis clúster
Número de clusters
2.5
5.0
7.5
10.0
12.5
lowBW
xDeath
AgeGroup2
RaceHispanic
xMarried
RaceBlack
xDrinking
xSmoking
AgeGroup1
RaceWhite
AgeGroup3
xSomeCollege
RaceAsian
RaceNative