



EXPLORANDO DATOS

La exploración es realmente importante para poder crear modelos de los datos.

Con la experiencia este proceso lo harás inconscientemente y conseguirás obtener mucha información de tus datos.

En esta hoja de trabajo pretendo darte una guía para poder aplicar la exploración paso a paso.

Utiliza las recetas del bloque 4 y del bloque 6 de exploración para lograrlo.

Vamos a explorar los datos de esta tabla:

`data(PimaIndiansDiabetes)`

Para ello sigo los pasos:

ENTIENDE TU TABLA DE DATOS

Entiende las variables de la tabla de datos.

	pregnant	glucose	pressure	triceps	insulin	mass	pedigree	age	diabetes
1	6	148	72	35	0	33.6	0.627	50	pos
2	1	85	66	29	0	26.6	0.351	31	neg
3	8	183	64	0	0	23.3	0.672	32	pos
4	1	89	66	23	94	28.1	0.167	21	neg
5	0	137	40	35	168	43.1	2.288	33	pos
6	5	116	74	0	0	25.6	0.201	30	neg
7	3	78	50	32	88	31.0	0.248	26	pos

Date unos segundos para entender las variables para entrar en el problema e interpretar los gráficos.

Nombre Variable	¿Qué característica mide?	¿Qué unidades tiene?	¿Qué tipo de variable es?
pregnant	Número de embarazos de la paciente	Cantidad de embarazos	Cuantitativa discreta
glucose	Concentración de glucosa en Sangre	mg/dL	Cuantitativa discreta
pressure	Presión diastólica en milímetros de mercurio	mm Hg	Cuantitativa discreta
triceps	Grosor del pliegue cutáneo del tríceps	mm	Cuantitativa discreta
insulin	Insulina sérica de 2 horas	mu U/ml	Cuantitativa discreta
mass	Índice de masa corporal	Peso en kg/altura en m2	Cuantitativa continua
pedigree	La función de pedigrí de diabetes calcula la probabilidad de diabetes según la edad del sujeto y sus antecedentes familiares diabéticos	probabilidad	Cuantitativa continua
age	Edad del paciente	años	Cuantitativa discreta
diabetes	Prueba de diabetes	factor	Cualitativa nominal

EXPLORANDO DATOS

DEFINE EL OBJETIVO DEL PROYECTO

Te lo defino yo de momento.

Por ejemplo:

“Queremos calcular un modelo clasificador que nos permite identificar si el paciente tiene diabetes o no utilizando los datos que tenemos.”

Con este objetivo en mente nos toca describir los datos.

RESUME LAS VARIABLES

Calcula los estadísticos de las variables.

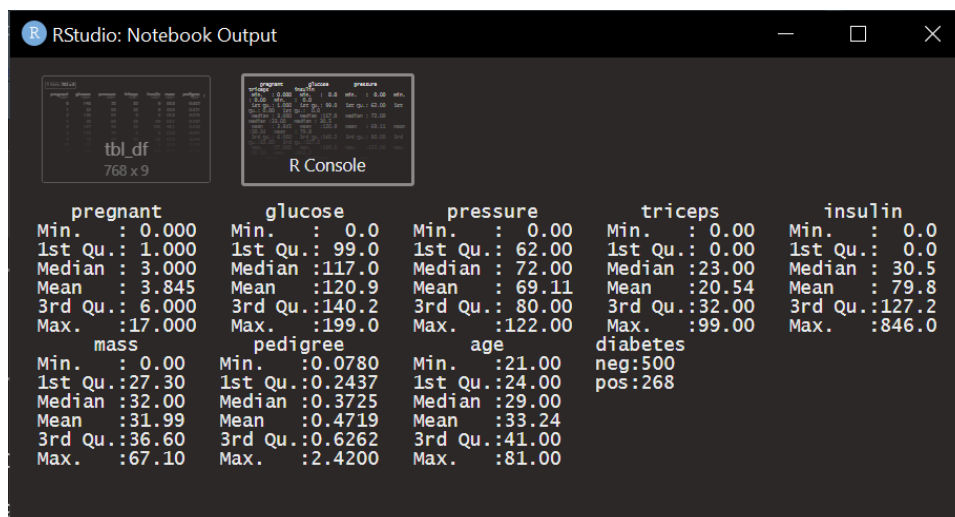
Si la variable es cuantitativa. Calcula:

- La media
- Desviación estándar
- Q1
- Q2
- Q3
- IQR
- IC de la media
- Error estándar (EE)
- Número de muestras (N)
- Kurtosis
- Asimetría

Si la variable es cualitativa. Calcula:

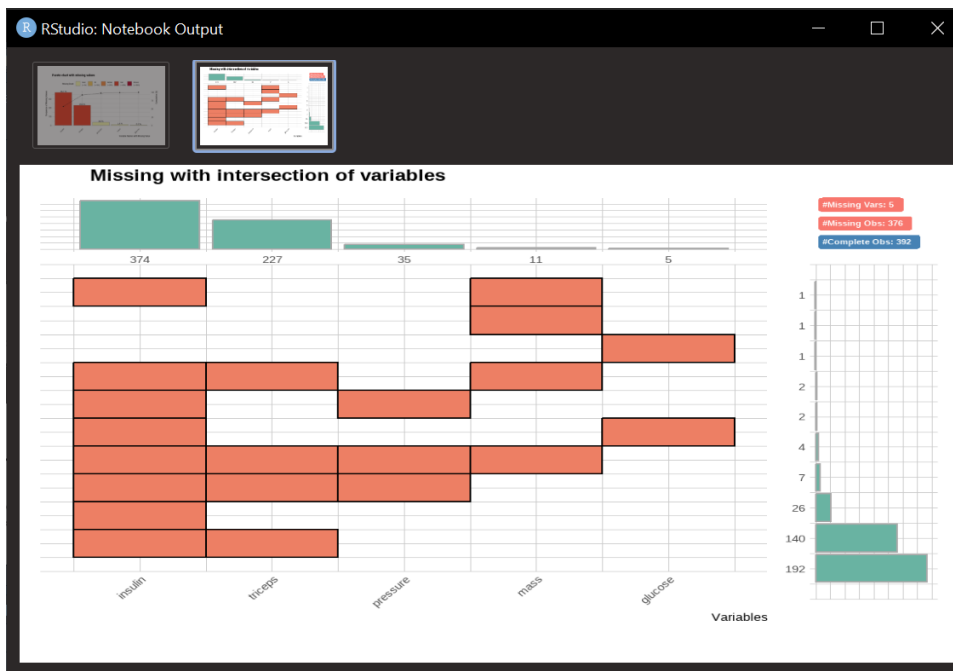
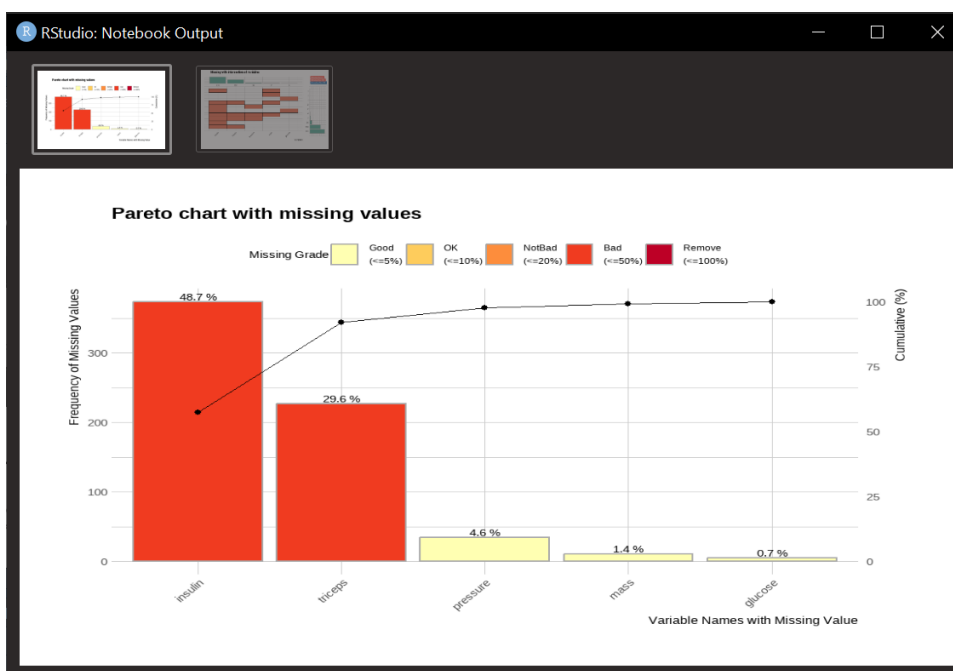
- Las frecuencias absolutas por grupo
- Las proporciones por grupo

Analisis inicial de datos



EXPLORANDO DATOS

Existen varios valores en cero que deben ser sustituidos para mejorar las estadísticas de los datos, pero se conserva la variable pregnant porque si puede ser cero.



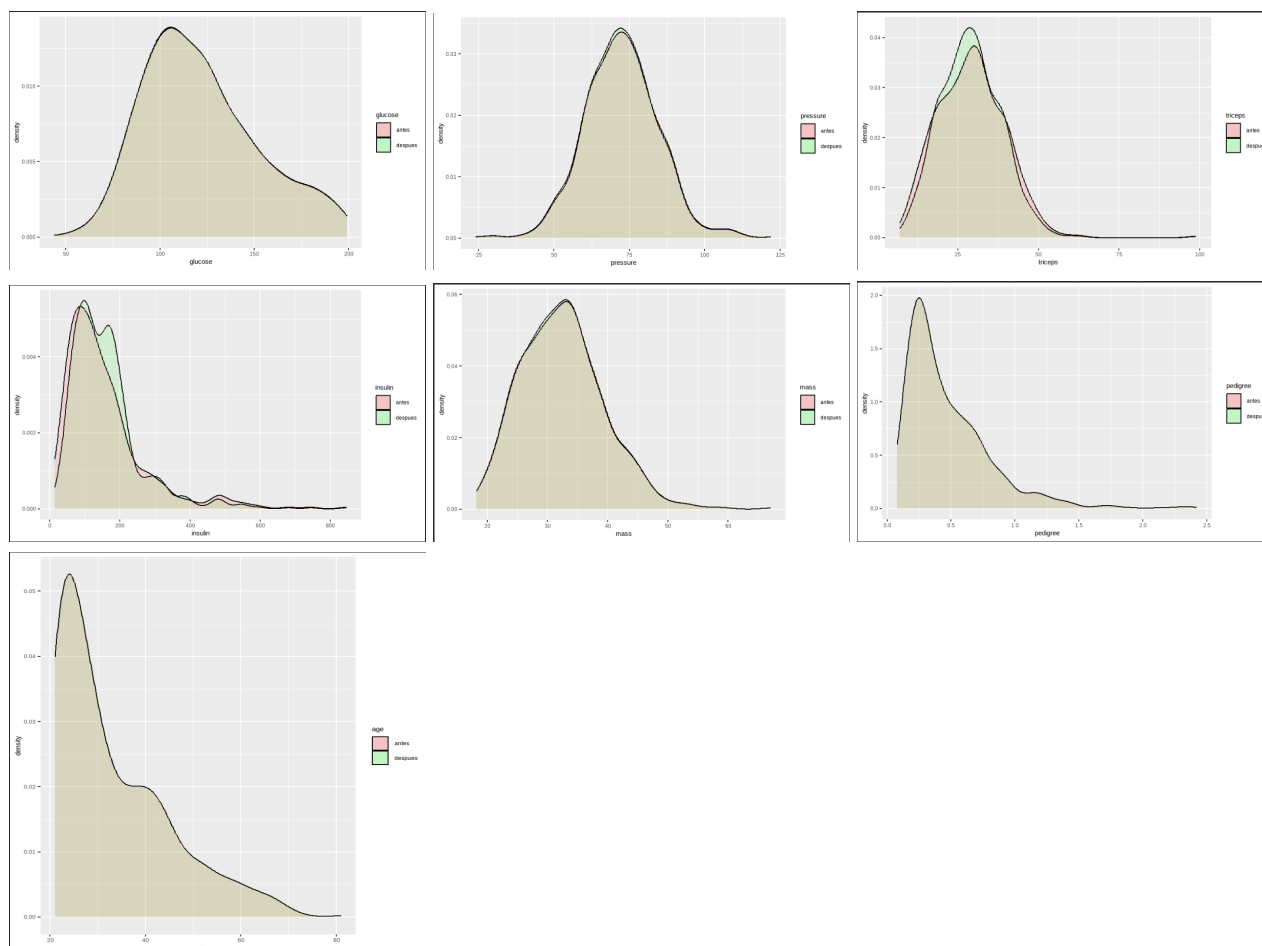
La variable insulin y tricep representan casi el 80% de los valores nulos. Únicamente el 51% de las filas están completas

EXPLORANDO DATOS

RStudio: Notebook Output

pregnant	glucose	pressure	triceps	insulin
Min. : 0.000	Min. : 44.0	Min. : 24.00	Min. : 7.00	Min. : 14.00
1st Qu.: 1.000	1st Qu.: 99.0	1st Qu.: 64.00	1st Qu.: 22.00	1st Qu.: 91.11
Median : 3.000	Median : 117.0	Median : 72.00	Median : 28.97	Median : 139.78
Mean : 3.845	Mean : 121.7	Mean : 72.39	Mean : 28.99	Mean : 158.03
3rd Qu.: 6.000	3rd Qu.: 141.0	3rd Qu.: 80.00	3rd Qu.: 35.55	3rd Qu.: 191.25
Max. : 17.000	Max. : 199.0	Max. : 122.00	Max. : 99.00	Max. : 846.00

mass	pedigree	age	diabetes
Min. : 18.20	Min. : 0.0780	Min. : 21.00	neg:500
1st Qu.: 27.50	1st Qu.: 0.2437	1st Qu.: 24.00	pos:268
Median : 32.25	Median : 0.3725	Median : 29.00	
Mean : 32.43	Mean : 0.4719	Mean : 33.24	
3rd Qu.: 36.60	3rd Qu.: 0.6262	3rd Qu.: 41.00	
Max. : 67.10	Max. : 2.4200	Max. : 81.00	



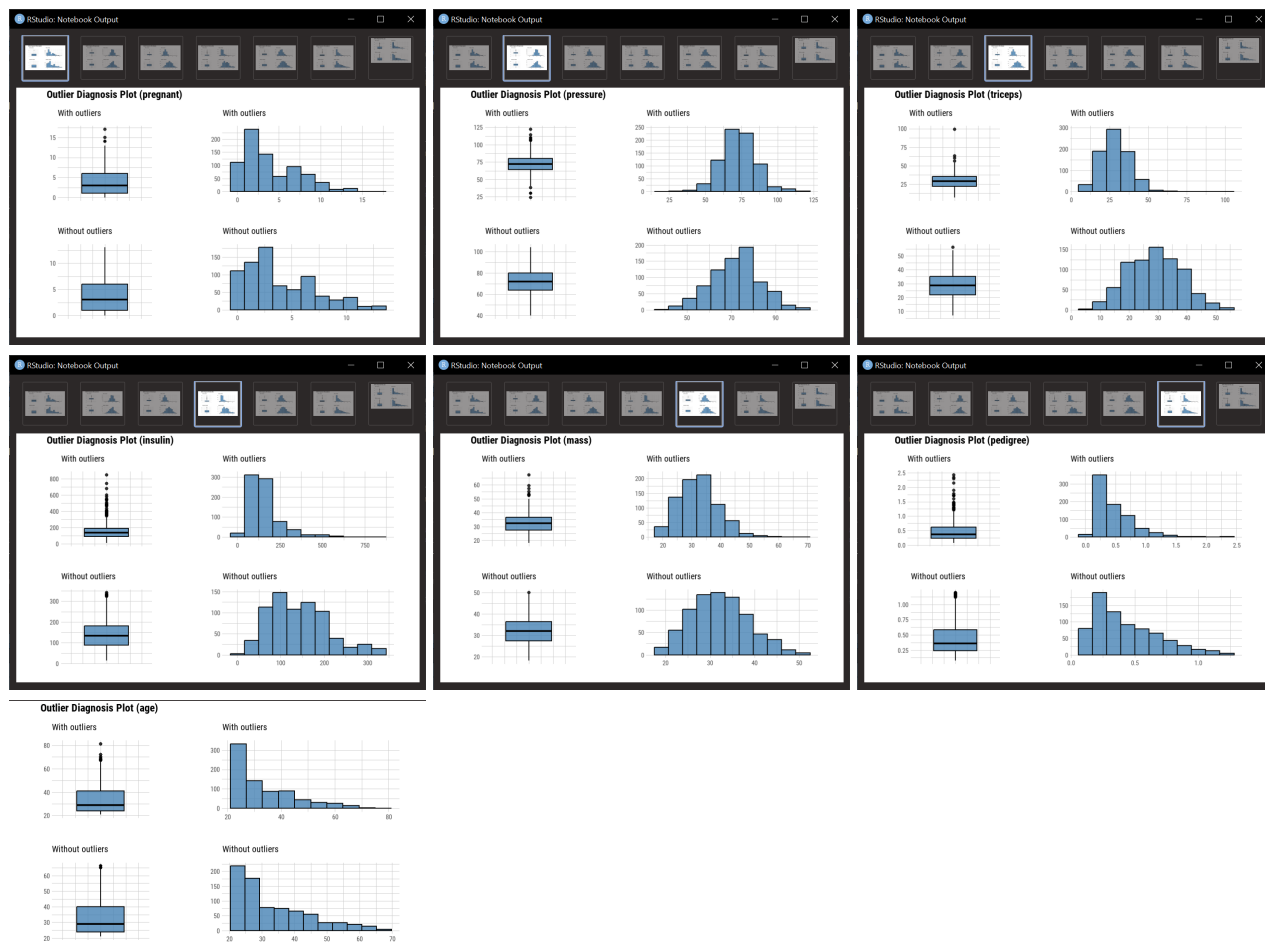
El resultado después de aplicar la función `missRanger` indica que aun la variable `insuline` puede requerir ajuste debido a outliers

Diagnóstico de outliers

EXPLORANDO DATOS

RStudio: Notebook Output

variables	outliers_cnt	outliers_ratio	outliers_mean	with_mean	without_mean
pregnant	4	0.5208333	15.000000	3.8450521	3.7866492
glucose	0	0.0000000		121.6516411	121.6516411
pressure	14	1.8229167	87.285714	72.3914112	72.1148591
triceps	3	0.3906250	74.000000	28.9903934	28.8138852
insulin	38	4.9479167	469.009931	158.0276858	141.8395690
mass	8	1.0416667	56.187500	32.4328729	32.1828242
pedigree	29	3.7760417	1.543276	0.4718763	0.4298322
age	9	1.1718750	70.000000	33.2408854	32.8050066



Se observa diferencias importantes entre dejar los outliers o eliminarlos. Con lo anterior es mejor seleccionar el método para imputar outliers que no afecte las estadísticas y también podemos aprovechar los datos existentes para inferir los datos en vez de eliminarlos y conservar la mayor cantidad de información para el modelo de clasificación.

Variables numéricas

EXPLORANDO DATOS

BLOQUE 6 - RStudio Source Editor

estadistica_descriptiva_imp

Filter

	pregnant	glucose	pressure	triceps	insulin	mass	pedigree	age
N	768.000	768.000	768.000	768.000	768.000	768.000	768.000	768.000
Media	3.845	121.652	72.401	28.935	141.737	32.437	0.428	33.241
Desv_Estandar	3.370	30.479	12.138	9.296	66.480	6.882	0.245	11.760
Mediana	3.000	117.000	72.000	28.970	139.665	32.250	0.372	29.000
IQR	5.000	42.000	16.000	13.553	88.936	9.100	0.338	17.000
Min	0.000	44.000	32.408	7.000	14.000	18.200	0.078	21.000
Max	17.000	199.000	110.000	58.796	340.821	56.608	1.191	81.000
Rango	17.000	155.000	77.592	51.796	326.821	38.408	1.113	60.000
Cuartil1	1.000	99.000	64.000	22.000	91.114	27.500	0.244	24.000
Cuartil3	6.000	141.000	80.000	35.553	180.050	36.600	0.582	41.000
Error_Estandar	0.122	1.100	0.438	0.335	2.399	0.248	0.009	0.424
IC95MediaLower	3.607	119.496	71.543	28.277	137.036	31.950	0.410	32.409
IC95MediaUpper	4.083	123.807	73.260	29.592	146.439	32.924	0.445	34.073
Varianza	11.354	928.954	147.335	86.418	4419.615	47.367	0.060	138.303
Suma	2953.000	93428.460	55604.236	22221.806	108854.396	24911.808	328.449	25529.000
Asimetria	0.900	0.532	0.157	0.251	0.678	0.569	0.954	1.127
Curtosis	3.150	2.724	3.727	2.858	3.231	3.574	3.283	3.631

Showing 1 to 17 of 17 entries, 8 total columns

De acuerdo a la descripción de las variables los valores mínimos de las variables glucose hasta mass no podrían ser cero.

Se observan variables con sesgo hacia la derecha ocasionados por outliers: pregnant, glucose, insulin, pedigree, y age.

De las anteriores la variable insulin tiene bastante dispersión con una varianza bastante elevada en comparación a las demás.

En segundo lugar, la variable glucose también presenta una dispersión grande.

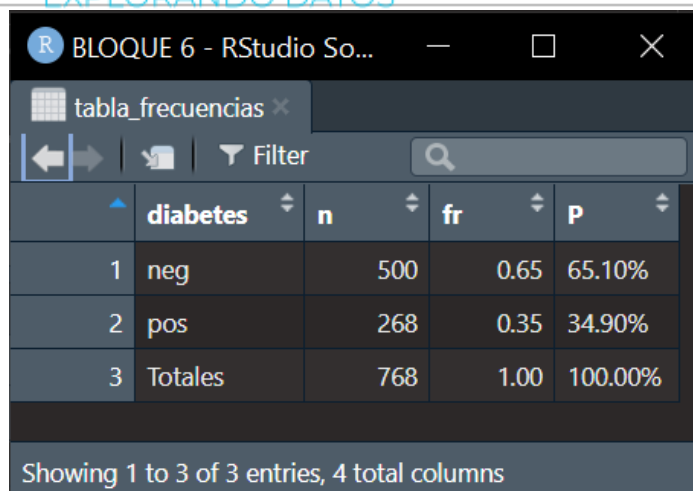
En ambos casos podemos suponer una campana achatada.

Las variables pressure y mass tiene asimetría izquierda su cola izquierda debería ser más alargada que la cola derecha. El resto de las variables lo hacen hacia el lado contrario.

Con lo anterior se intuye que se requiere hacer una limpieza de datos para poder comparar los escenarios solo de datos completos, sin outliers o con valores imputados.

Variables cualitativas

EXPLORANDO DATOS



The screenshot shows an RStudio window titled 'BLOQUE 6 - RStudio So...'. A data frame named 'tabla_frecuencias' is open, displaying a frequency table for the variable 'diabetes'. The table has four columns: 'diabetes', 'n', 'fr', and 'p'. It contains three rows: 'neg' (500 cases, 0.65 frequency, 65.10% proportion), 'pos' (268 cases, 0.35 frequency, 34.90% proportion), and 'Totales' (768 cases, 1.00 frequency, 100.00% proportion). The status bar at the bottom indicates 'Showing 1 to 3 of 3 entries, 4 total columns'.

	diabetes	n	fr	p
1	neg	500	0.65	65.10%
2	pos	268	0.35	34.90%
3	Totales	768	1.00	100.00%

En la variable cualitativa diabetes el 65% de los casos son negativos, se podría requerir balancear los datos.

DISTRIBUCIONES UNIVARIADAS

Dibuja las distribuciones de todas las variables cuantitativas por separado.

Identifica:

- Forma de la distribución

EXPLORANDO DATOS

- Simetría
- Si se parece a una distribución normal
- Outliers
- Valores centrales y dispersión

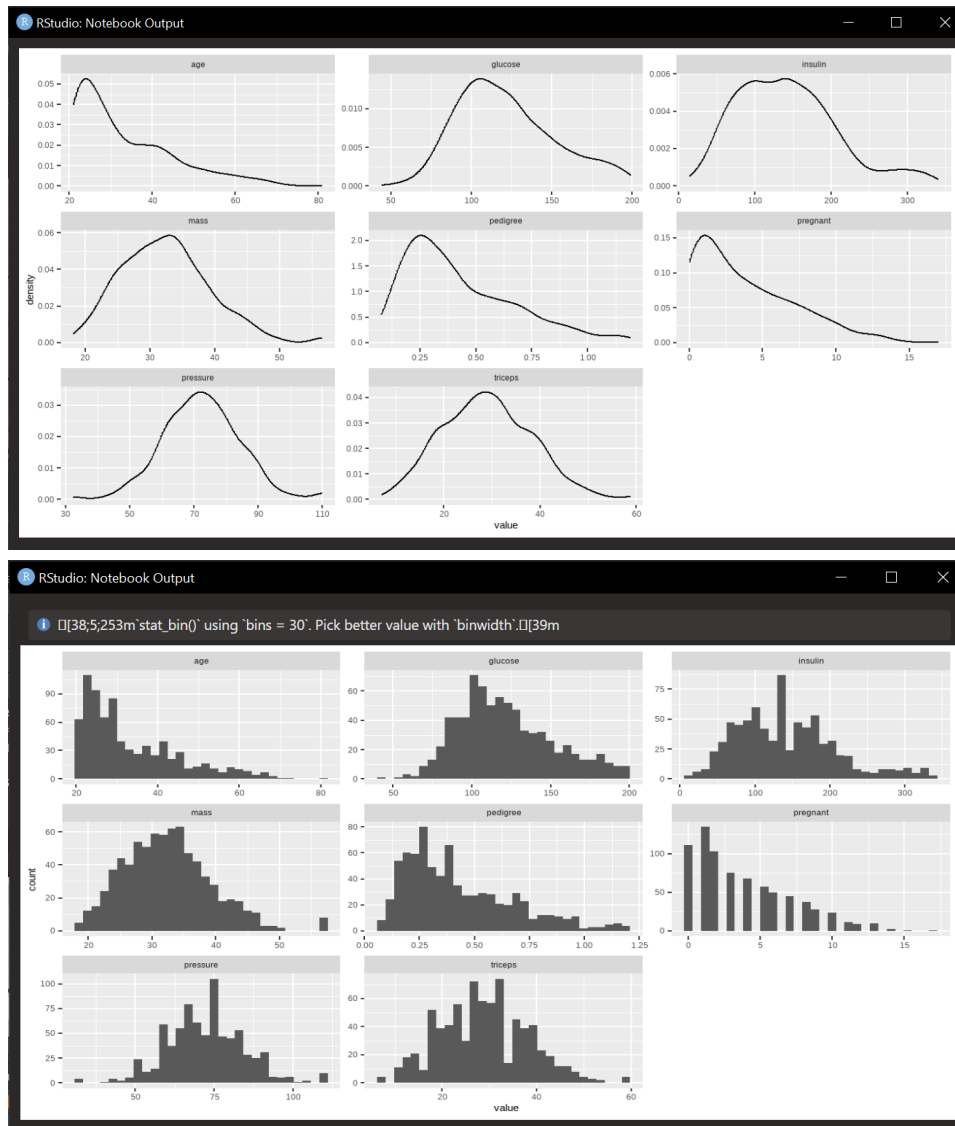
Utiliza:

- Histograma de densidad
- Boxplots
- Qqplots

Para las variables cualitativas, utiliza:

- El diagrama de barras o sectores

Distribuciones de frecuencia

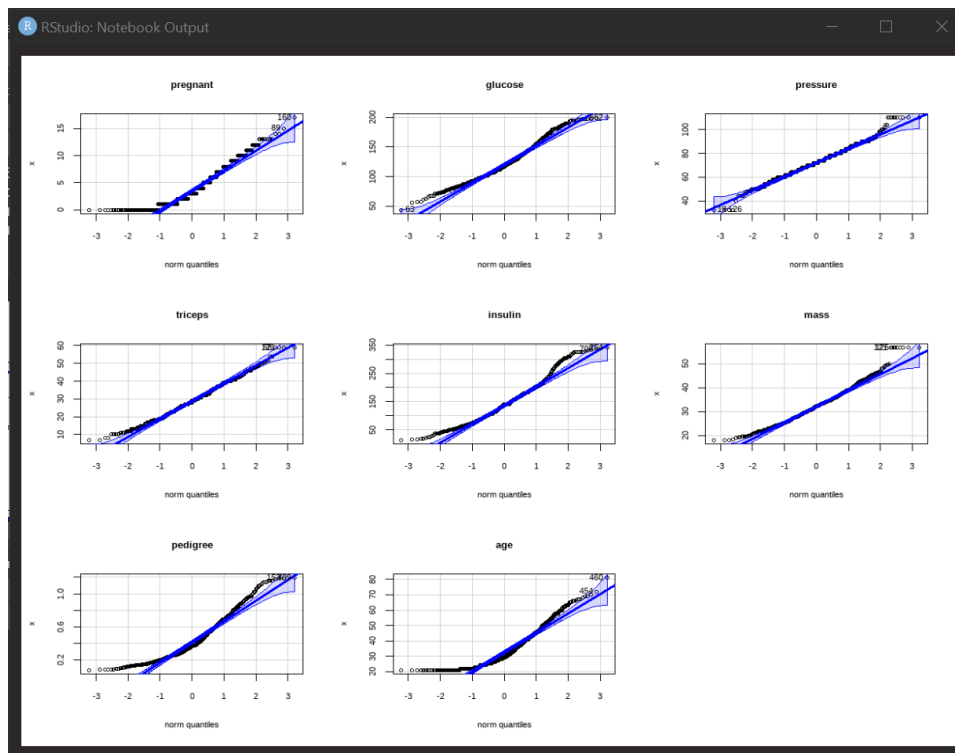


EXPLORANDO DATOS

Variables con apariencia de distribución normal: glucose, pressure, mass, pedigree, insulin y triceps.

Se observa simetría en las variables pressure, mass y triceps

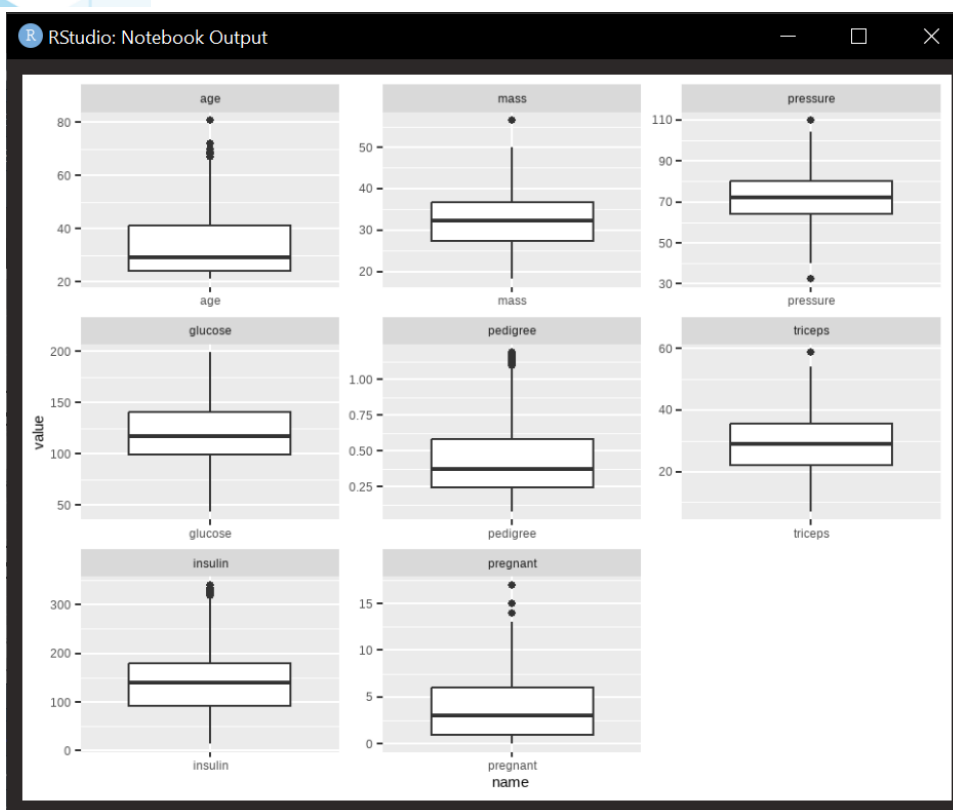
Variables con apariencia de distribución logarítmica: pregnant, age.



Gráficamente con el qqplot ninguna de las variables se ajusta a una distribución normal.

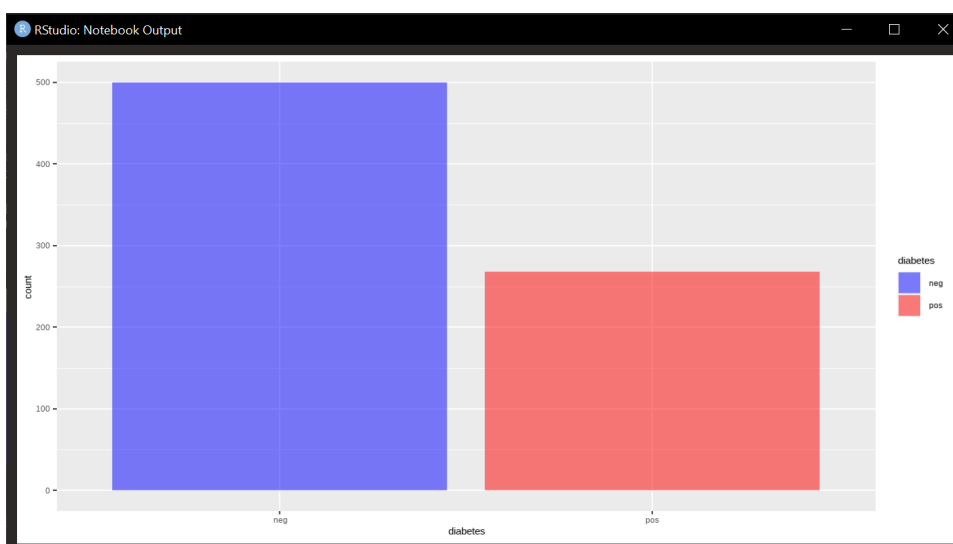
Outliers:

EXPLORANDO DATOS



Se observa que las variables age, pedigree y pregnant presentan varios outliers si los comparamos a las demás variables, es estos casos específicos, las cajas se ven desplazadas hacia arriba o hacia abajo lo que indican que no están muy bien centradas.

Diagrama de barras para la variable diabetes



EXPLORANDO DATOS

EXPLORANDO DATOS DISTRIBUCIONES Y RESUMEN MULTIVARIADO

Vamos a hacer lo mismo que has hecho anteriormente pero por grupos de la variable respuesta.

En este caso, la diabetes.

Dibuja las distribuciones y calcula los resúmenes numéricos por cada grupos de diabetes yes, no)

Grupo con diabetes positivo

BLOQUE 6 - RStudio Source Editor

estadistica_descriptiva

Filter

	pregnant	glucose	pressure	triceps	insulin	mass	pedigree	age
N	268.000	268.000	268.000	268.000	268.000	268.000	268.000	268.000
Media	4.866	141.257	70.825	22.164	100.336	35.143	0.550	37.067
Desv_Estandar	3.741	31.940	21.492	17.680	138.689	7.263	0.372	10.968
Mediana	4.000	140.000	74.000	27.000	0.000	34.250	0.449	36.000
IQR	6.250	48.000	16.000	36.000	167.250	7.975	0.465	16.000
Min	0.000	0.000	0.000	0.000	0.000	0.000	0.088	21.000
Max	17.000	199.000	114.000	99.000	846.000	67.100	2.420	70.000
Rango	17.000	199.000	114.000	99.000	846.000	67.100	2.332	49.000
Cuartil1	1.750	119.000	66.000	0.000	0.000	30.800	0.262	28.000
Cuartil3	8.000	167.000	82.000	36.000	167.250	38.775	0.728	44.000
Error_Estandar	0.229	1.951	1.313	1.080	8.472	0.444	0.023	0.670
IC95MediaLower	4.418	137.433	68.251	20.047	83.731	34.273	0.506	35.754
IC95MediaUpper	5.314	145.081	73.398	24.281	116.941	36.012	0.595	38.380
Varianza	13.997	1020.139	461.898	312.572	19234.673	52.751	0.139	120.303
Suma	1304.000	37857.000	18981.000	5940.000	26890.000	9418.200	147.534	9934.000
Asimetria	0.501	-0.493	-1.933	0.115	1.833	0.001	1.713	0.578
Curtosis	2.544	4.383	7.590	2.772	7.257	7.653	7.452	2.636

Showing 1 to 17 of 17 entries, 8 total columns

Grupo con diabetes negativo



EXPLORANDO DATOS

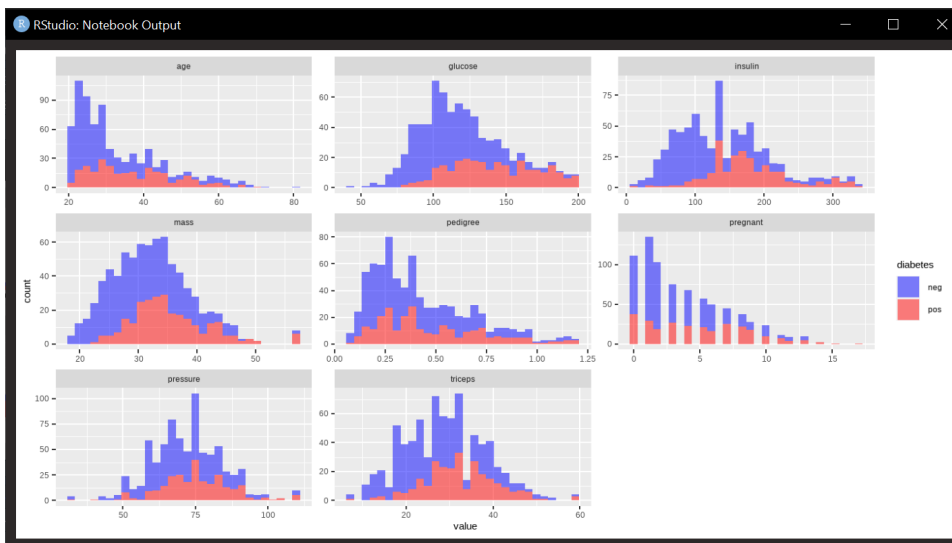
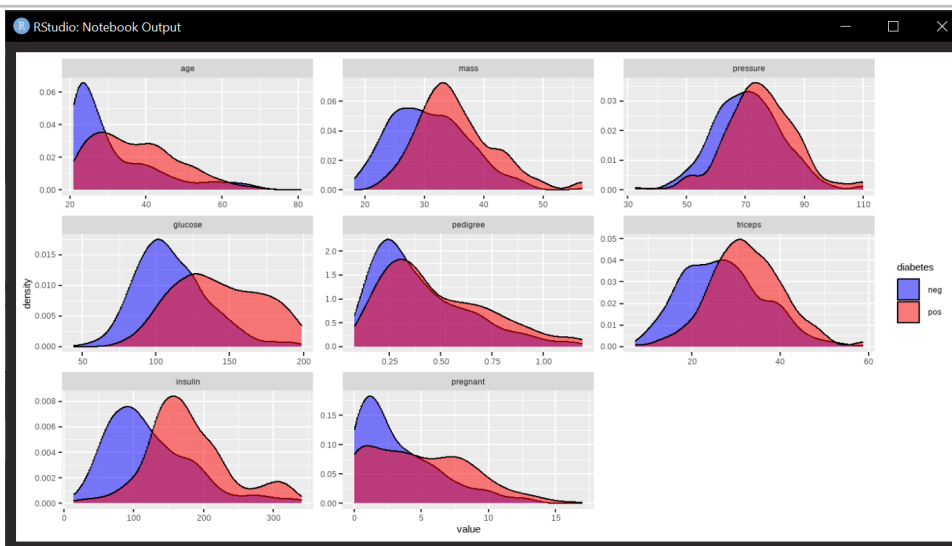
BLOQUE 6 - RStudio Source Editor

estadistica_descriptiva

Filter

	pregnant	glucose	pressure	triceps	insulin	mass	pedigree	age
N	500.000	500.000	500.000	500.000	500.000	500.000	500.000	500.000
Media	3.298	109.980	68.184	19.664	68.792	30.304	0.430	31.190
Desv_Estandar	3.017	26.141	18.063	14.890	98.865	7.690	0.299	11.668
Mediana	2.000	107.000	70.000	21.000	39.000	30.050	0.336	27.000
IQR	4.000	32.000	16.000	31.000	105.000	9.900	0.332	14.000
Min	0.000	0.000	0.000	0.000	0.000	0.000	0.078	21.000
Max	13.000	197.000	122.000	60.000	744.000	57.300	2.329	81.000
Rango	13.000	197.000	122.000	60.000	744.000	57.300	2.251	60.000
Cuartil1	1.000	93.000	62.000	0.000	0.000	25.400	0.230	23.000
Cuartil3	5.000	125.000	78.000	31.000	105.000	35.300	0.562	37.000
Error_Estandar	0.135	1.169	0.808	0.666	4.421	0.344	0.013	0.522
IC95MediaLower	3.034	107.689	66.601	18.359	60.126	29.630	0.404	30.167
IC95MediaUpper	3.562	112.271	69.767	20.969	77.458	30.978	0.456	32.213
Varianza	9.103	683.362	326.275	221.711	9774.345	59.134	0.089	136.134
Suma	1649.000	54990.000	34092.000	9832.000	34396.000	15152.100	214.867	15595.000
Asimetria	1.111	0.173	-1.804	0.031	2.491	-0.664	2.000	1.567
Curtosis	3.668	4.870	8.618	1.960	12.351	6.016	9.048	4.932

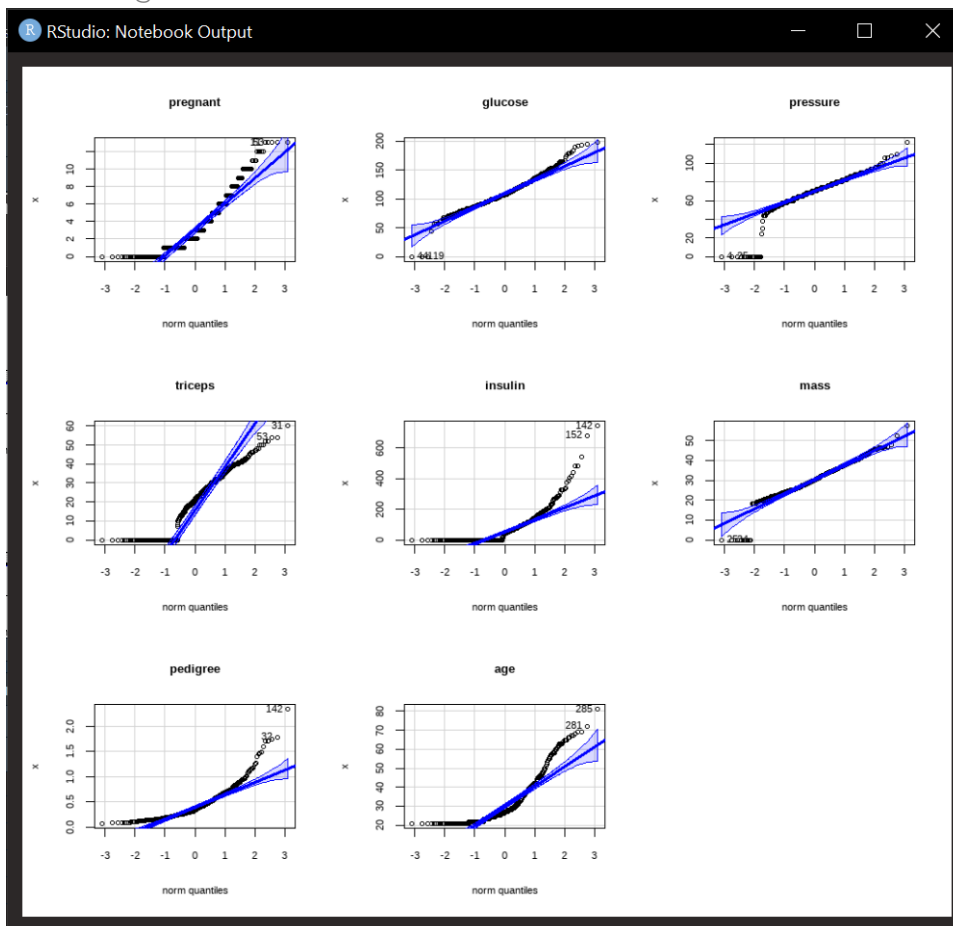
Showing 1 to 17 of 17 entries, 8 total columns



EXPLORANDO DATOS

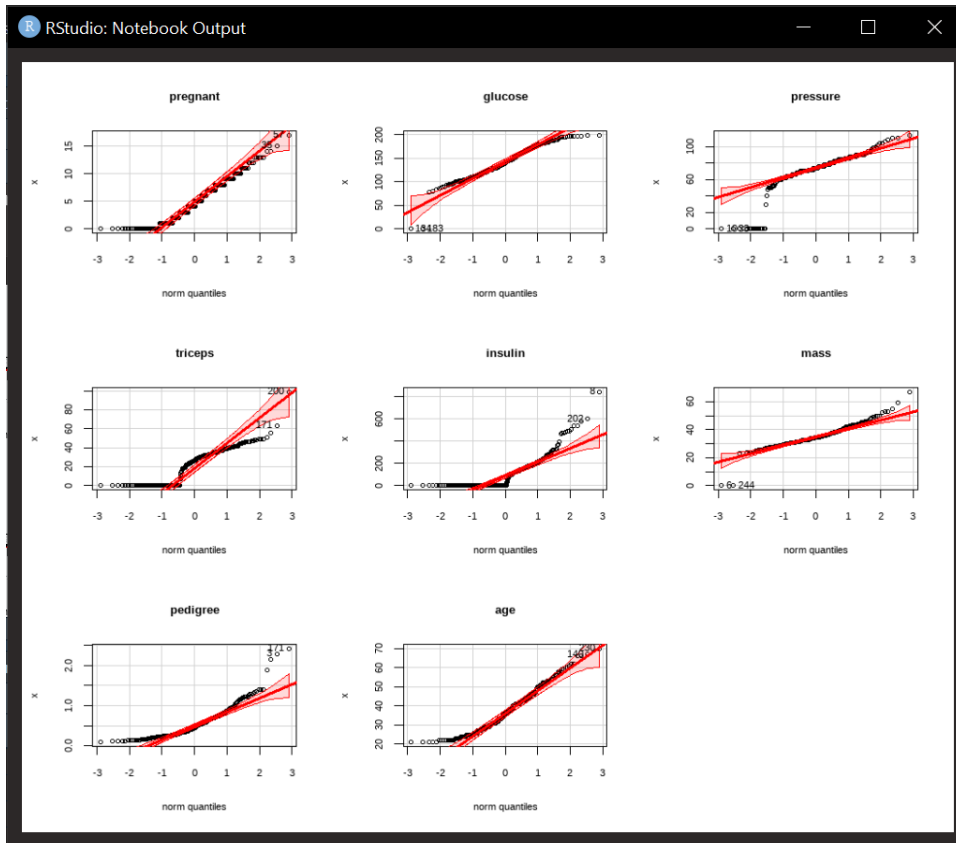


QQplot de casos negativos



QQplot de casos positivos

EXPLORANDO DATOS



EXPLORANDO DATOS

EXPLORANDO DATOS

CORRELACIÓN ENTRE VARIABLES

Calcula la correlación entre variables cuantitativas siguiendo:

- Matriz de correlación
- Matrixplot
- Matrixplot con coeficientes de correlación
- Diagrama de correlación
- Matrixplot por grupos de diabetes

Identifica relaciones entre las variables y si se ven entre grupos.

Matriz de correlación

BLOQUE 6 - RStudio Source Editor

matriz_correlacion

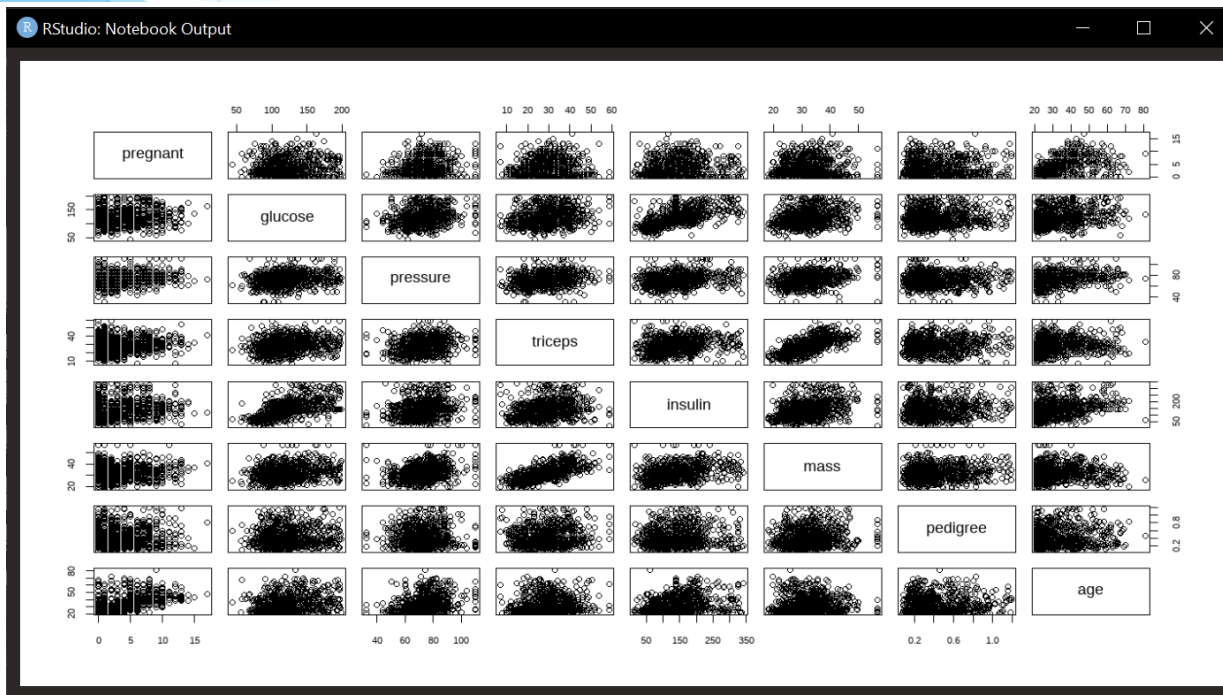
Filter

	pregnant	glucose	pressure	triceps	insulin	mass	pedigree	age
pregnant	1.000	0.130	0.218	0.118	0.143	0.023	0.004	0.544
glucose	0.130	1.000	0.234	0.232	0.608	0.236	0.083	0.269
pressure	0.218	0.234	1.000	0.257	0.232	0.300	0.021	0.336
triceps	0.118	0.232	0.257	1.000	0.256	0.705	0.088	0.149
insulin	0.143	0.608	0.232	0.256	1.000	0.316	0.109	0.318
mass	0.023	0.236	0.300	0.705	0.316	1.000	0.128	0.025
pedigree	0.004	0.083	0.021	0.088	0.109	0.128	1.000	0.050
age	0.544	0.269	0.336	0.149	0.318	0.025	0.050	1.000

Showing 1 to 8 of 8 entries, 8 total columns

Matrixplot

EXPLORANDO DATOS



Matrixplot con coeficientes de correlación

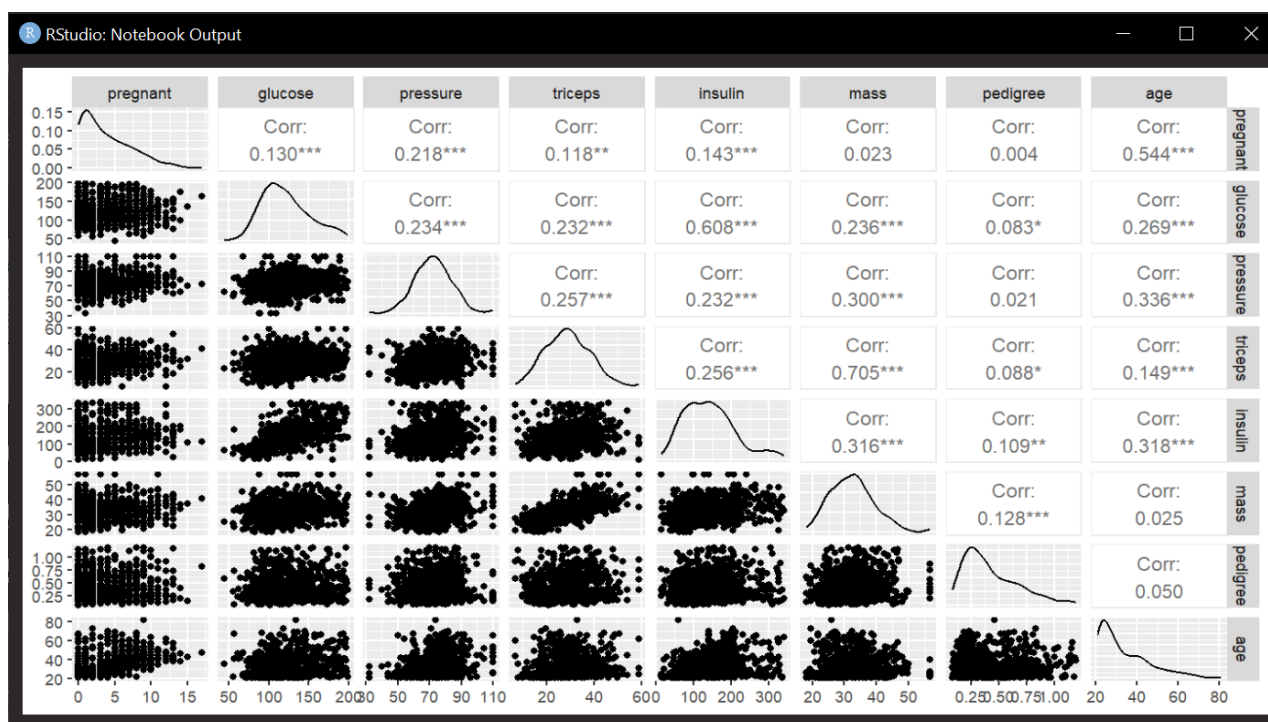
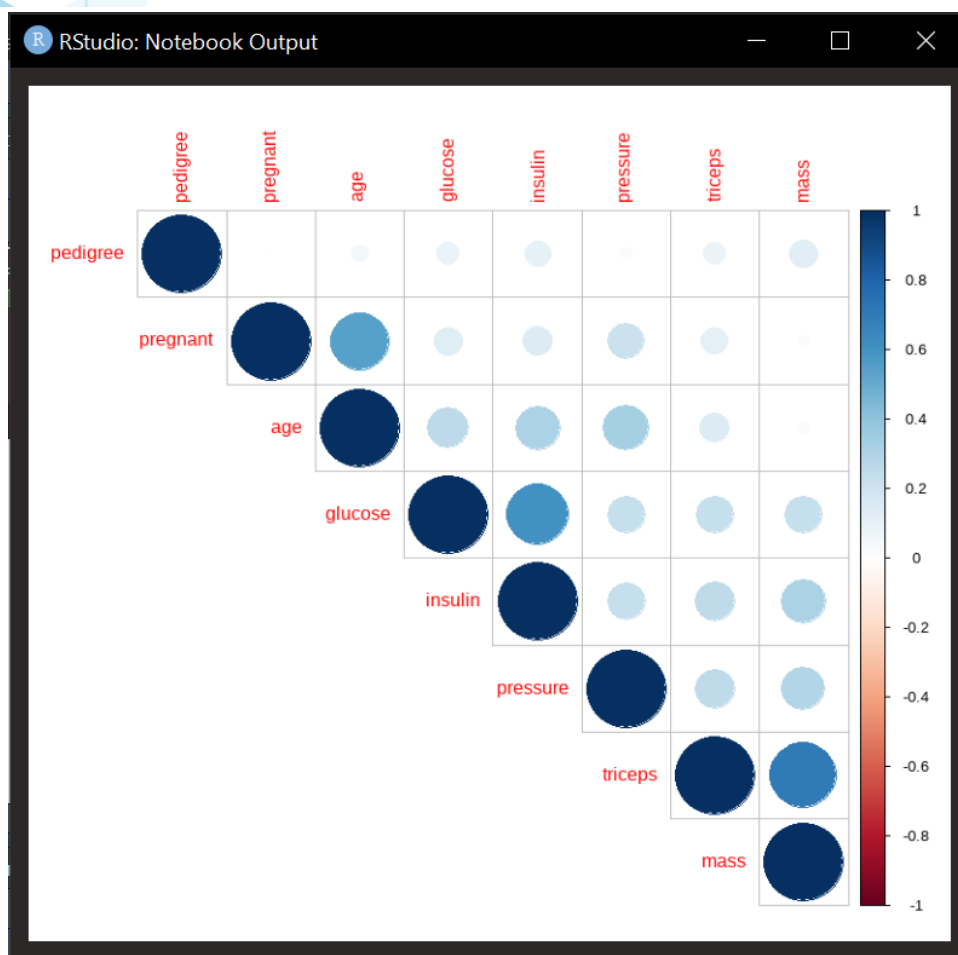
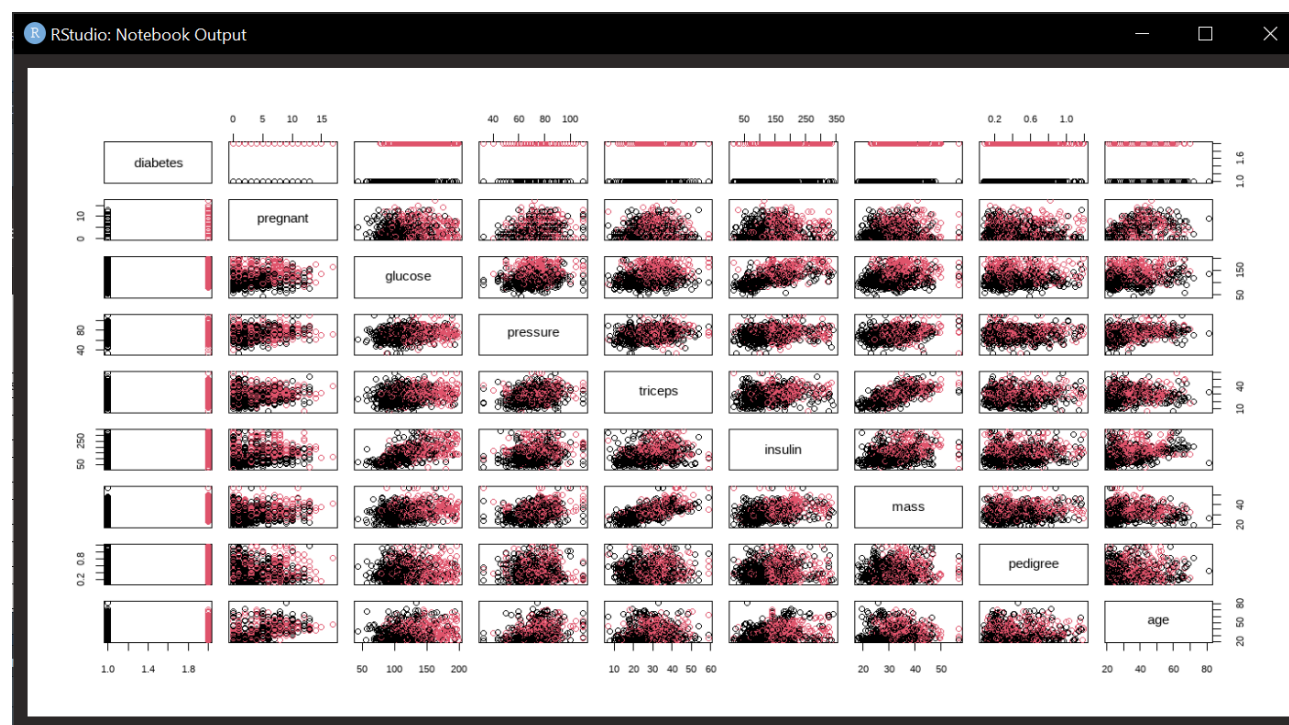


Diagrama de correlación

EXPLORANDO DATOS

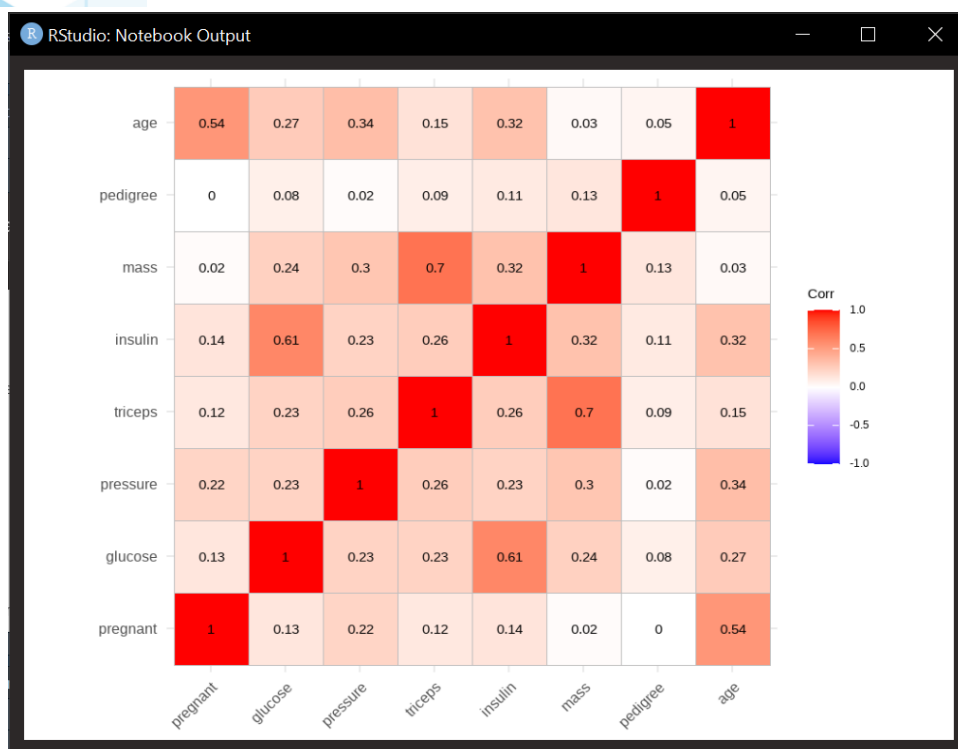


Matrixplot por grupos de diabetes



Mapa de calor

EXPLORANDO DATOS



EXPLORANDO DATOS

Ahora, describe lo más relevante que has visto.

Siempre teniendo en cuenta el objetivo que te has marcado al principio.

¡A por ello!

EXPLORANDO DATOS

Conclusiones de la exploración:

Muchos de los datos requirieron de limpieza para poder realizar el análisis exploratorio de manera más precisa. Se limpiaron valores no razonables y outliers principalmente en las variables insulin y triceps que representan casi el 80% de la muestra de datos incluso en el cruce de estas mismas variables. Antes que eliminar se prefiere conservar el volumen de datos con sustituciones realizadas.

Se observan correlaciones fuertes entre pares de variables:
glucose-insulin, mass-triceps, age-pregnant.

En general, esto se comporta de manera lógica ya que se esperan esas relaciones estrechas entre azúcar en sangre y niveles altos de insulina, a mayor masa muscular mayores triceps, entre más edad más partos han tenido. Esto podría representar redundancia de información para el modelo de clasificación.

La variable Edad tiene una fuerte relación con todas las demás variables en dos sentidos:

Entre mayor en edad es el sujeto de estudio tiende a tener mayor presión arterial, mayores niveles de azúcar en sangre, mayor cantidad de insulina, menor índice de masa muscular y triceps.

Las variables pregnant y pedigree parecen no ser muy relevantes para determinar si el sujeto de estudio tiene diabetes.

Se puede considerar el entrenamiento del modelo únicamente con los datos completos y contrastarlo con el presente volumen de datos imputado para ver la precisión del mismo.

¡Bien bien bien!

¡Este ejercicio es fundamental!

Te ayudará muchísimo en tus análisis ☺



EXPLORANDO DATOS