Heart_Disease EDA

Dataset :

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
1	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
2	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
3	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
4	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
5	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0
6	58	0	0	100	248	0	0	122	0	1.0	1	0	2	1
7	58	1	0	114	318	0	2	140	0	4.4	0	3	1	0
8	55	1	0	160	289	0	0	145	1	0.8	1	1	3	0
9	46	1	0	120	249	0	0	144	0	0.8	2	0	3	0
10	54	1	0	122	286	0	0	116	1	3.2	1	2	2	0
11	71	0	0	112	149	0	1	125	0	1.6	1	0	2	1
12	43	0	0	132	341	1	0	136	1	3.0	1	0	3	0
13	34	0	1	118	210	0	1	192	0	0.7	2	0	2	1
14	51	1	0	140	298	0	1	122	1	4.2	1	3	3	0
15	52	1	0	128	204	1	1	156	1	1.0	1	0	0	0
16	34	0	1	118	210	0	1	192	0	0.7	2	0	2	1
17	51	0	2	140	308	0	0	142	0	1.5	2	1	2	1
18	54	1	0	124	266	0	0	109	1	2.2	1	1	3	0
19	50	0	1	120	244	0	1	162	0	1.1	2	0	2	1
20	58	1	2	140	211	1	0	165	0	0.0	2	0	2	1
21	60	1	2	140	185	0	0	155	0	3.0	1	0	2	0

Number of columns and rows :

```
1025   14
```

Meaning of each variable:

Age= age of the patient (Continuous numerical)

Sex= gender of the patient (1 = male; 0 = female) (binary)

Cp = chest pain type (nominal categorical)

Trestbps= resting blood pressure (Continuous numerical)

Chol= cholesterol in mg/dl (Continuous numerical)

Fbs= fasting blood sugar >120 mg/dl (1 = true; 0 = false) (binary)

Restecg = electrocardiographic results (values: 0, 1, 2) the values represent The intensity and seriousness of the situation 2 is more intense than 0 or 1

(ordinal categorical)

Thalach= maximum heart rate achieved “” (Continuous numerical)

Exang= from working out the person had an enigna(pain in the chest because lack of oxygen) ? (1 = yes; 0 = no) (binary)

Oldpeak = It represents the depression of the ST segment exam observed in an electrocardiogram when the patient exercises,It is a numerical value that indicates how many millimeters the segment drops, and it is associated with possible issues in blood flow to the heart.

If the ST segment is depressed during exercise, it could indicate a problem with the coronary arteries. 1 (Continuous numerical)

oldpeak = 0.0 → no depression = normal

oldpeak = 1.5 → moderate depression

oldpeak = 3.0 → high depression = higher risk

slope = it’s the type of slope the patient had during the st segment exam (0,1,2)

(ordinal categorical)

2= “upsloping” = it’s the normal slope it doesn’t indicate seriousness or severity

1= “flatsloping”=it represents a moderate level of seriousness and severity

0=”downsloping”= it represents a higher level of seriousness and severity

Ca= This variable refers to the number of main coronary vassels (0-3) that are affected" (ordinal categorical)

Thal=this varibale refers to the type of defect the main vassels have (0-3)

(ordinal categorical)

1= normal

2= fixed defect

3= reversable defect

Target = it refers to the presence of heart disease in the patient. (0=no ,1=yes) (binary)

OBJECTIVE OF THE ANALYSIS

En este análisis voy a realizar un análisis exploratorio de datos para observar y entender cada variable del conjunto de datos, y comprender la correlación y las relaciones que cada una de esas variables tiene con la variable objetivo binaria, que indica si el paciente tiene o no enfermedad cardíaca (0,1).

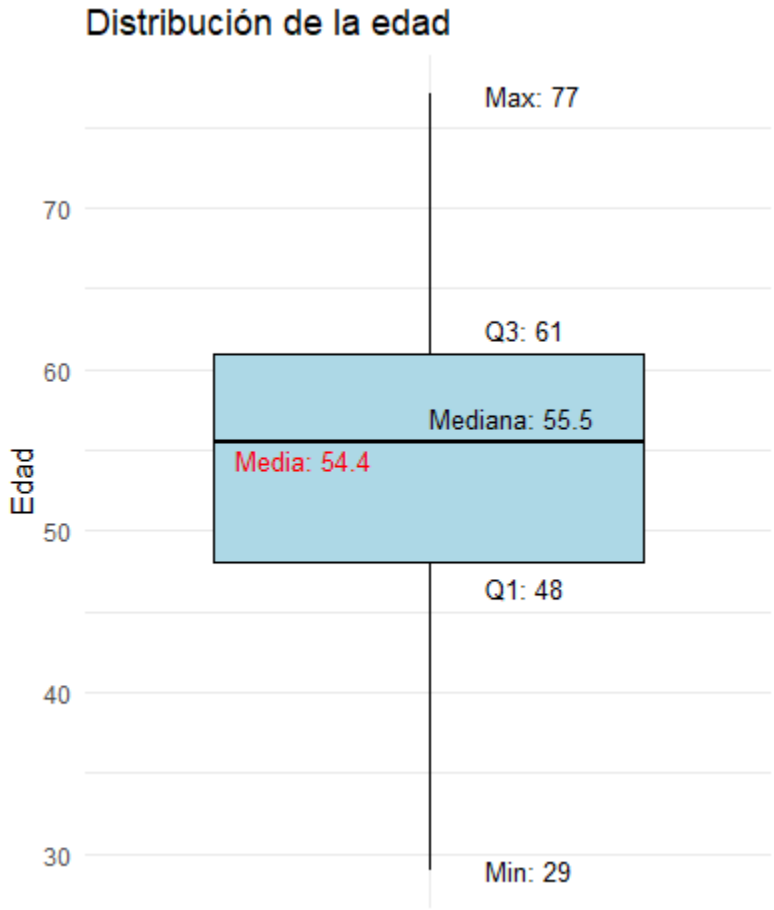
Missing// duplicate values

No hay valores faltantes en el conjunto de datos, pero este conjunto de datos tiene muchas filas duplicadas, en total 723 filas de 1025 filas. Por lo tanto, la fuente de donde se obtuvo este conjunto de datos no es confiable. Tal vez hubo un error durante la entrada de datos. Es imposible que 14 variables de salud diferentes relacionadas con la enfermedad cardíaca sean iguales tantas veces, así que decidí eliminar esas filas duplicadas.

```
> sum(duplicated(heart_disease))
[1] 723
> sum(is.na(heart_disease))
[1] 0
```

UNIVARIATE ANALYSIS :

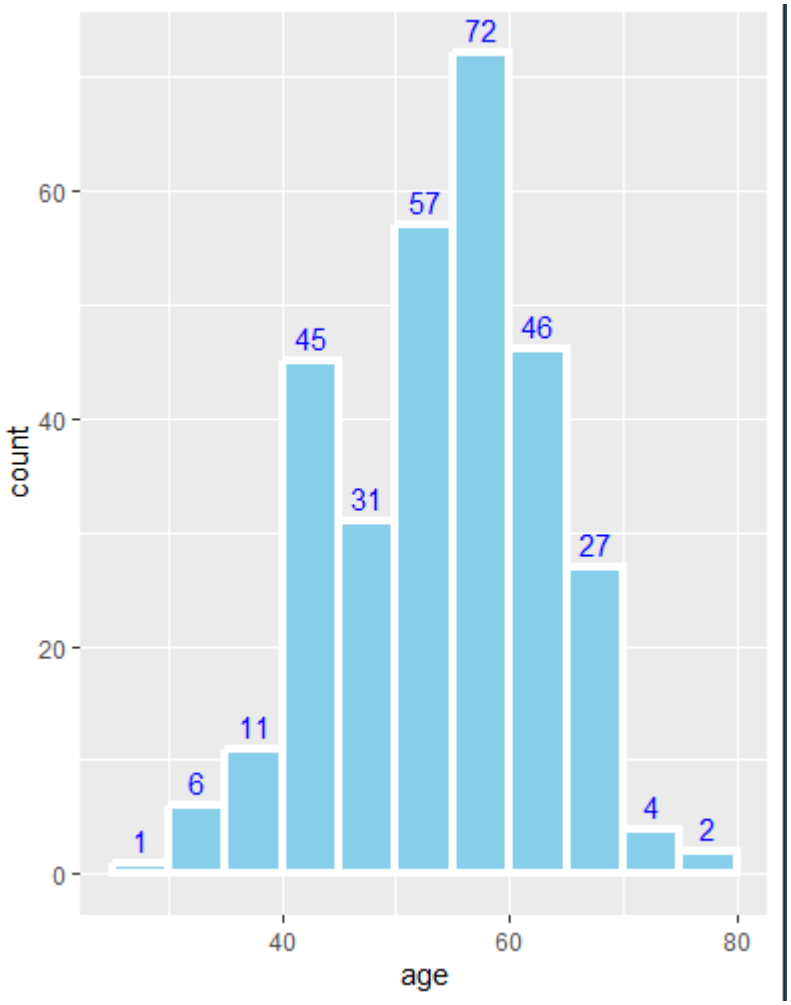
Age =




```
> sd(heart_disease$age)
[1] 9.07229
```

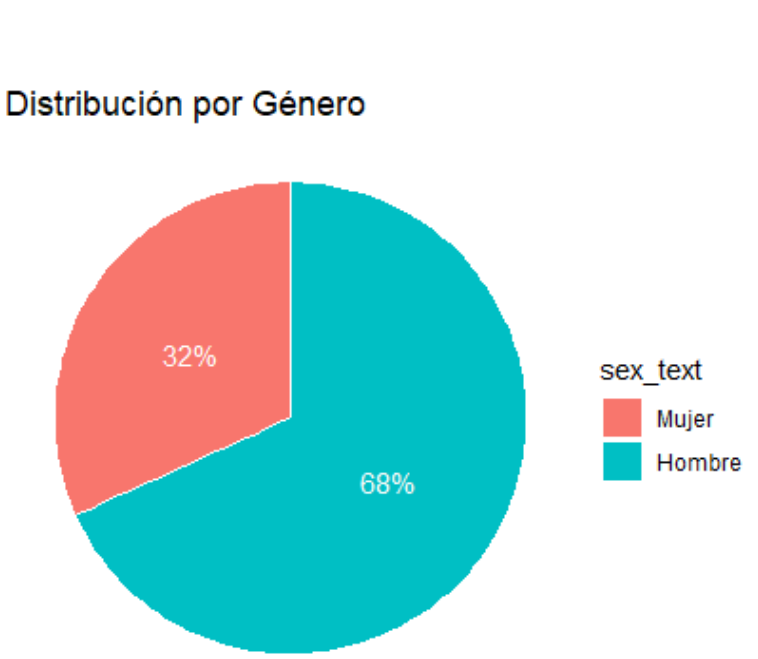
Su valor mínimo nos muestra que el paciente más joven en el conjunto de datos tiene 29 años y el paciente más viejo tiene 77 años. La media de edad es de 54 años. El 75 % de los pacientes tienen menos de 61 años, mientras que el 25 % restante tiene más de 61 años hasta llegar a los 77 años. El primer 25 % de los pacientes tienen edades entre 29 y 48 años.

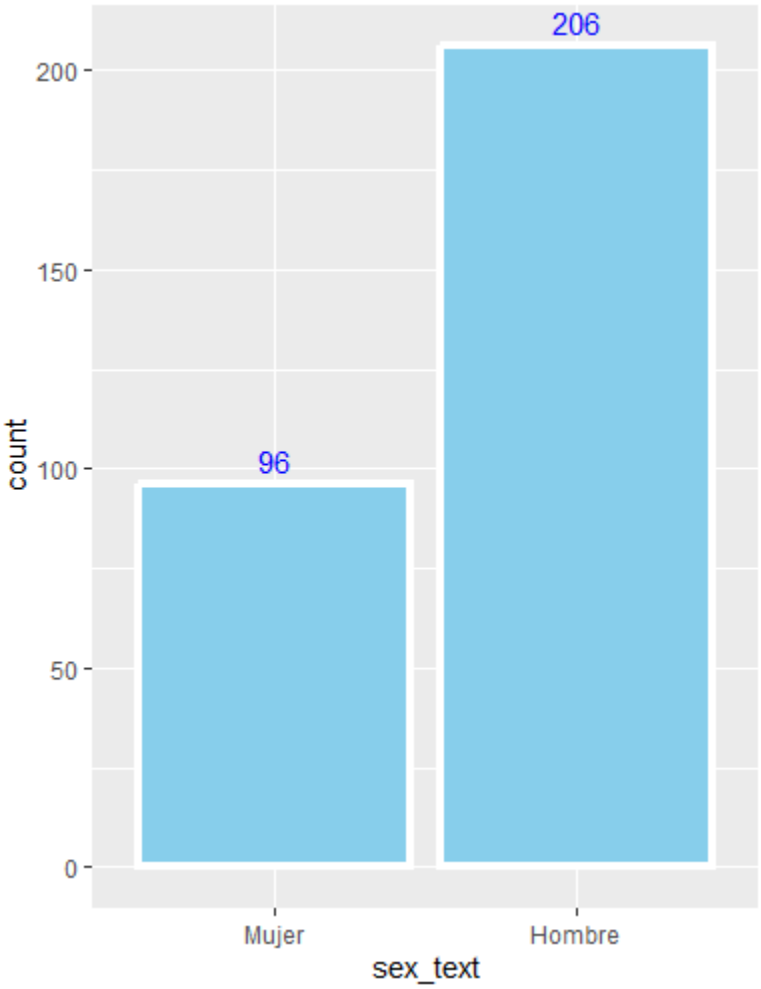
El 50 % de los pacientes tienen 56 años o más, y el otro 50 % tiene menos de 56 años. La edad de cada paciente varía alrededor de la media por ±9 años.



Aca podemos ver que la mayoria de edad de las personas que se les tomo este examen tienen entre 55-60 años con una frecuencia de 72 se va bajando gradualmente la frecuencia de los pacientes cuando la edad es menor a 40 años o cuando la edad es mayor a 65 el intervalo de 50 a 60 años es el intervalo que mas frecuencia hubo de pacientes con un total de 175 pacientes

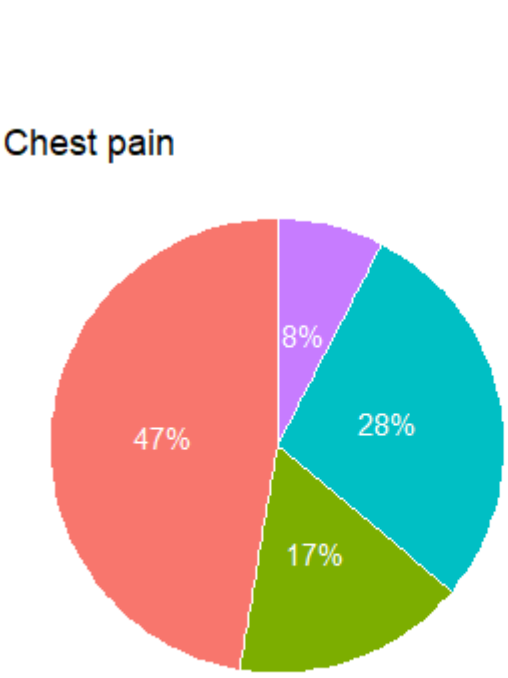
Sex:

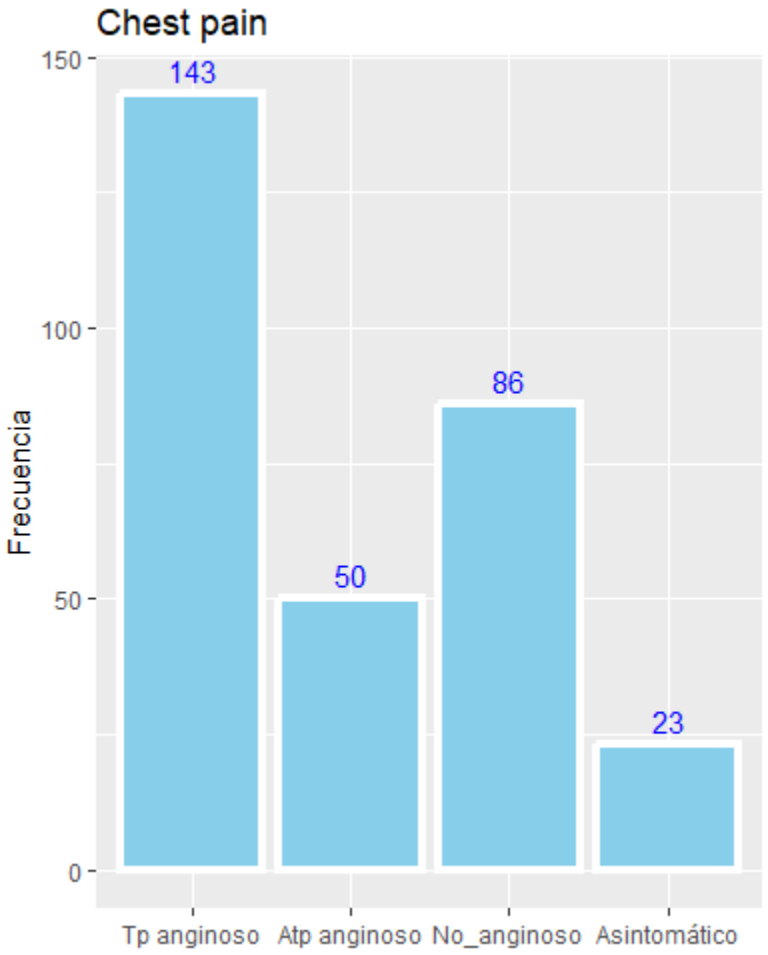




La variable genero despues de la limpieza quedo con una division de 32 % mujeres Vs 68 % hombres y una frecuencia en mujeres de 96 Vs 206 hombres

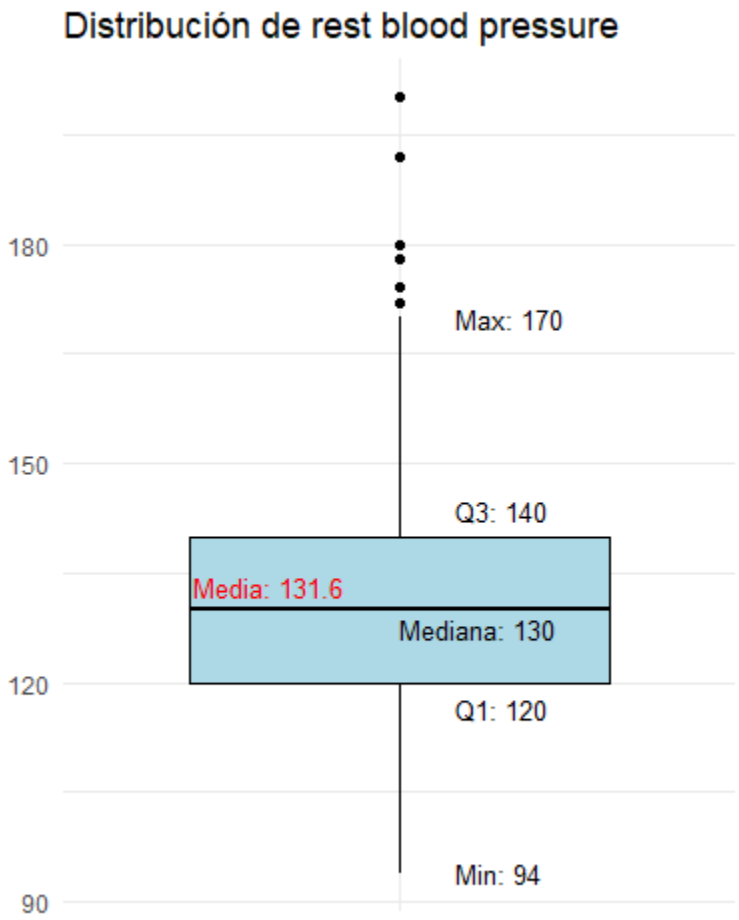
Chest pain:





La variable chest pain tiene una division de tp anginoso de 47 % o una frecuencia de 143 casos despues el segundo tipo de dolor de pecho con mas casos fue “no_anginoso” con 28 % o una frecuencia de 86 el siguiente es atp anginoso con una frecuencia de 17 % o una frecuencia de 50 ya por ultimo el tipo de dolor de pecho con menos casos fue asintomatico con un 8% o una frecuencia de 23.

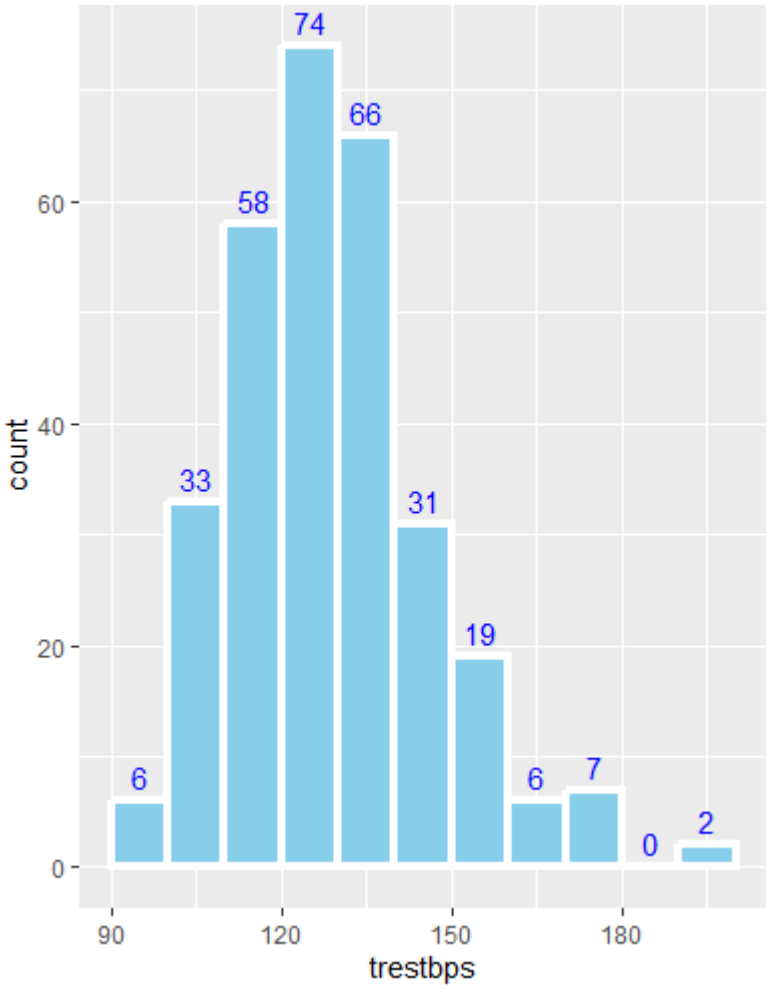
resting blood pressure:




```
> cat("standard deviation",sd(heart_disease$trestbps))
standard deviation 17.56339
```

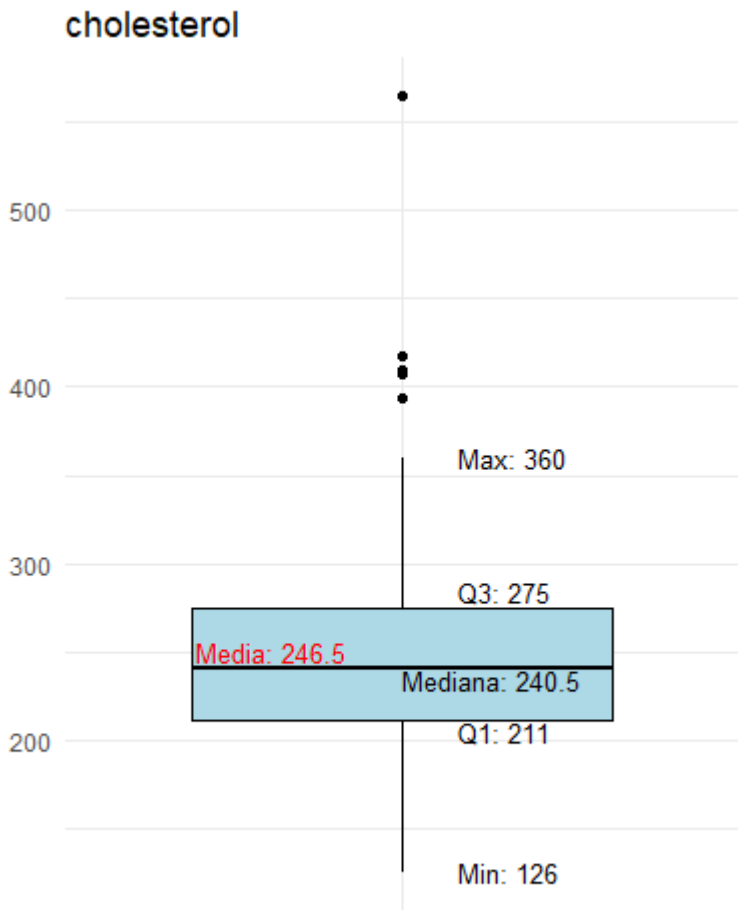
El minimo valor es 94 osea dentro de los pacientes este fue la minima presion sanguinea en reposo registrada y la maxima 170 el 75 % de los pacientes estan entre 94 de presion hasta 140 la mediana es 130 osea que el 50 % de los pacientes tienen mas de esa presion sanguinea y el otro 50

% menos de esa presion sanguinea la media de presion sanguinea de los pacientes fue 131 y la variabilidad es 17 osea con respecto a la media la presion sanguinea de los pacientes registrados varia 17± los outiliers parten apartir de 170 de presion sanguinea .



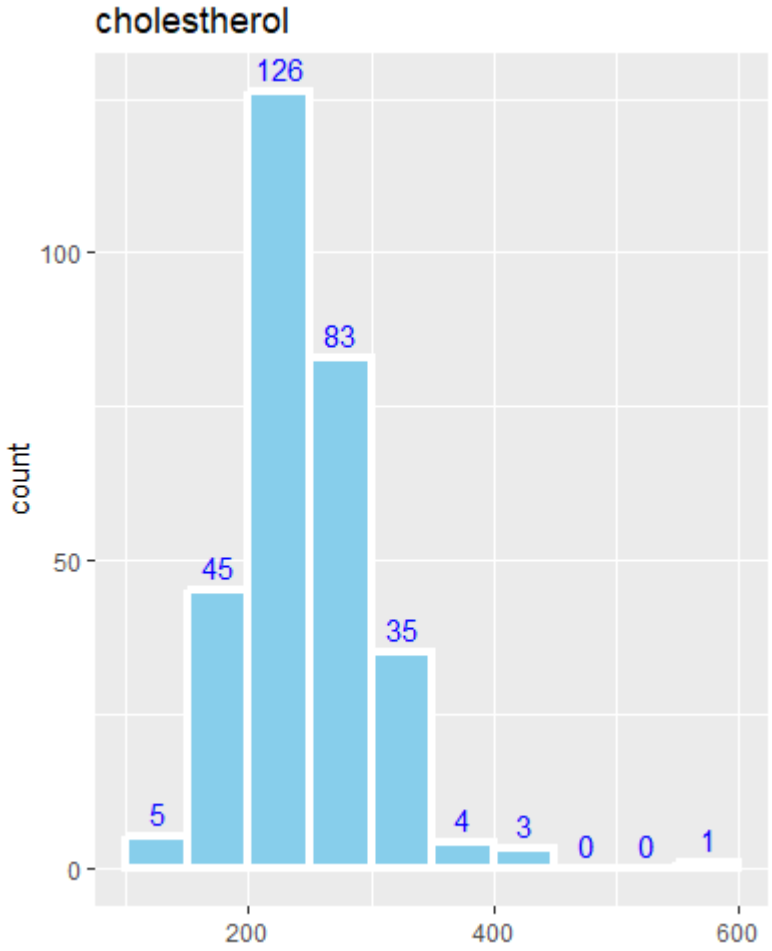
Como podemos ver el valor de rest blood pressure que mas se repite esta en el rango de 120-130 con frecuencia de 74 se puede decir que la mayoria de pacientes tienen entre 110 a 140 de blood pressure con una frecuencia de 198 pacientes se puede ver como apartir de 140 blood pressure empieza a bajar la frecuencia gradualmente

Cholestherol:



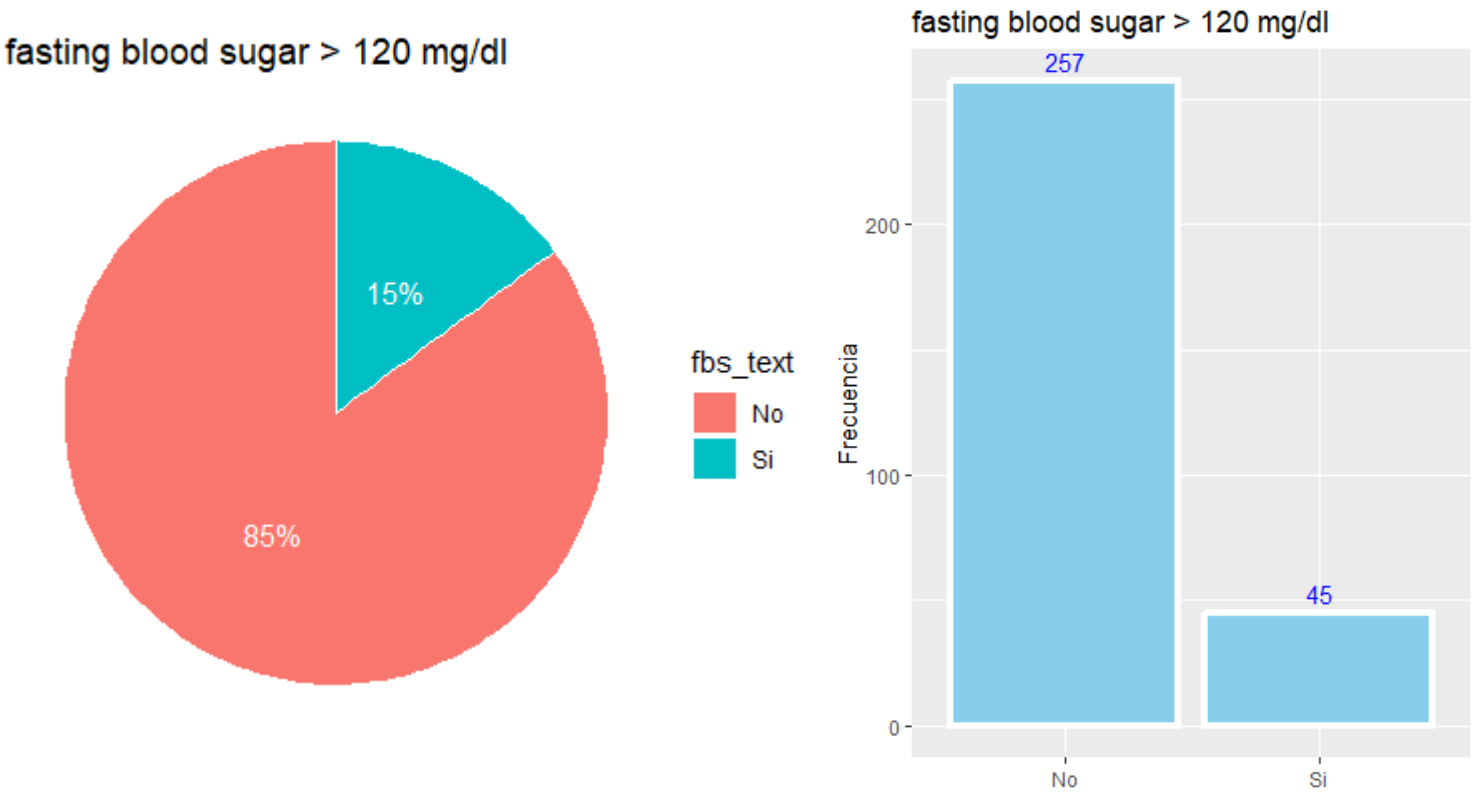

```
> cat("standard deviation",sd(heart_disease$chol))
standard deviation 51.75349
```

Los niveles de cholesterol promedio de los pacientes esta en 246, el 50 % de de los pacientes tiene entre 211 – 275 mg/dl de cholesterol el paciente con el min de cholesterol registrado fue 126 y el maximo 360 “hay mas de 360 pero son outliers “ un 25 % de los pacientes esta entre 126-211 y el otro 25 %esta entre 275 a 360 la varianza promedio en cuanto la media en esta variable es de 51.75 ±



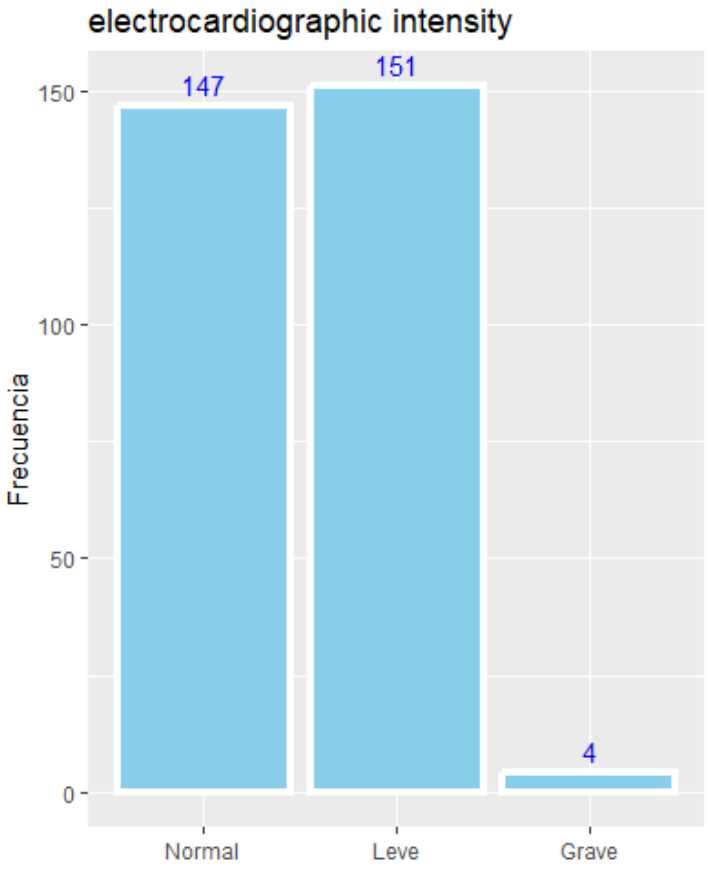
200 mg/dl de cholesterol fue el dato con mas frecuencia con 126 se puede ver que entre 150-250 mg/dl de cholesterol hay un total de 254 pacientes que significaria que entre ese intervalo esta el 84% de los pacientes totales apartir de 450 se estarian los outliers

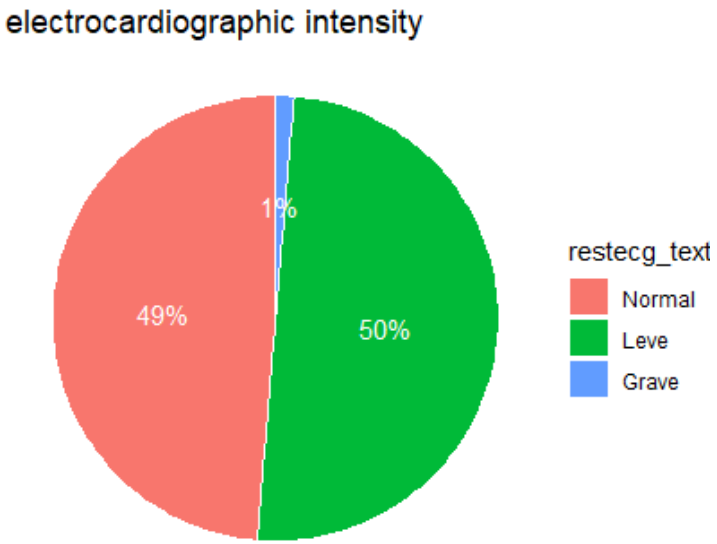
Fasting blood sugar:



El 85 % de los pacientes no tienen en ayunas un nivel de azucar mayor a 120 pero restante 15 % si esto en frecuencia equivaldria a no=257 , si = 45 pacientes

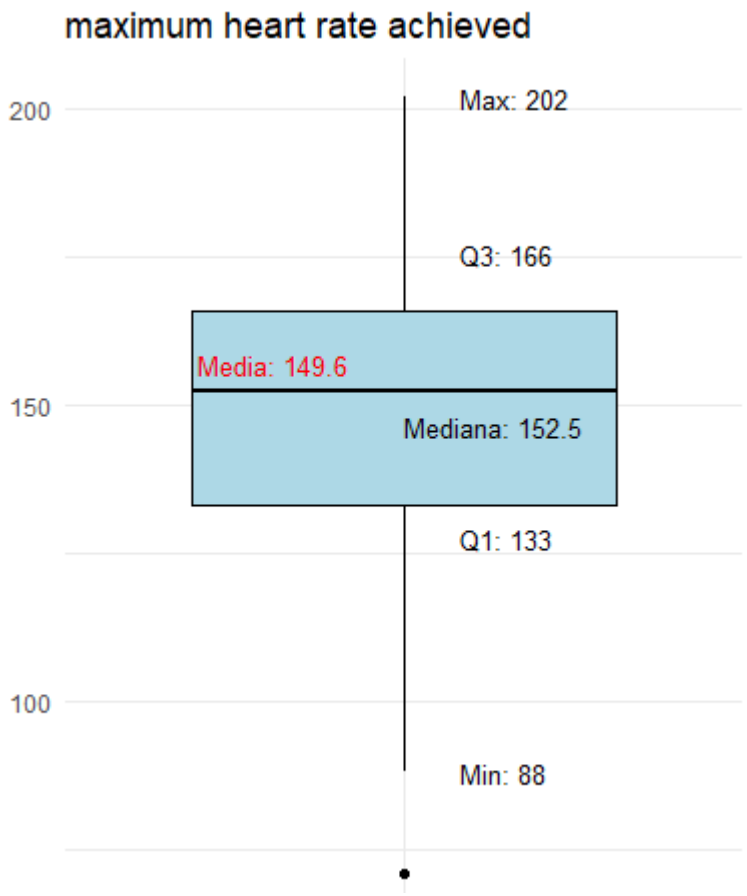
electrocardiographic results:





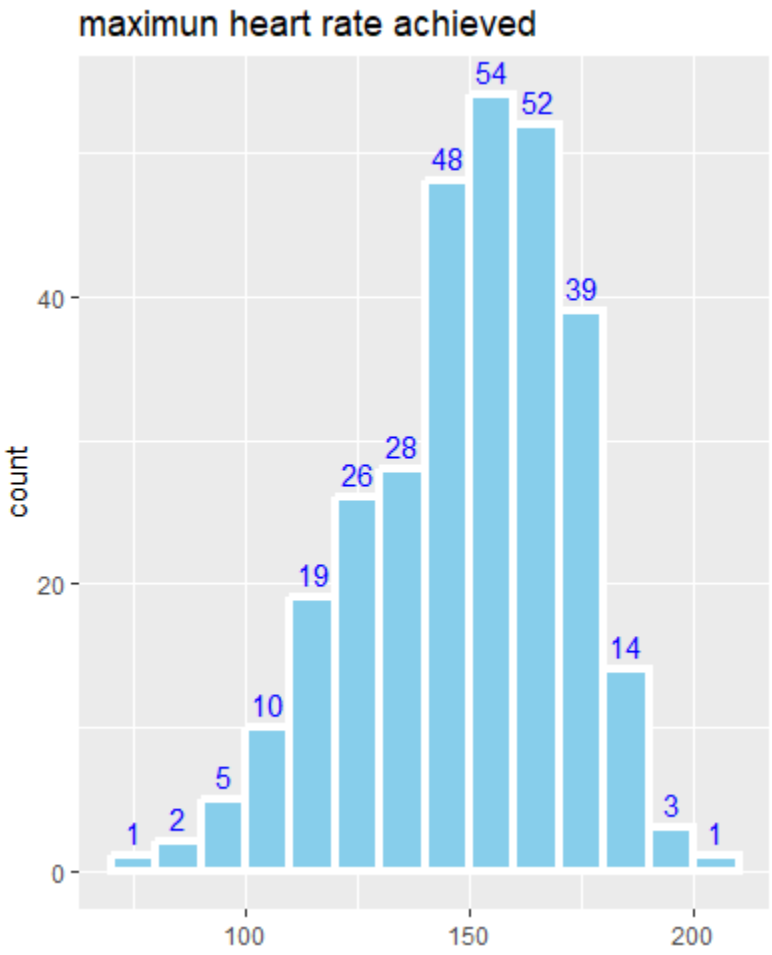
El 99 % de los pacientes tuvieron un resultado en el electrocardiograma de normal-leve solo 1 % obtuvo un resultado grave,esto en frecuencia equivaldria a que en normal-leve estuvieron 298 pacientes y solo 4 pacientes obtuvieron un mal diagnostico en el diagnostico del electrocardiograma (ECG)

maximum heart rate achieved



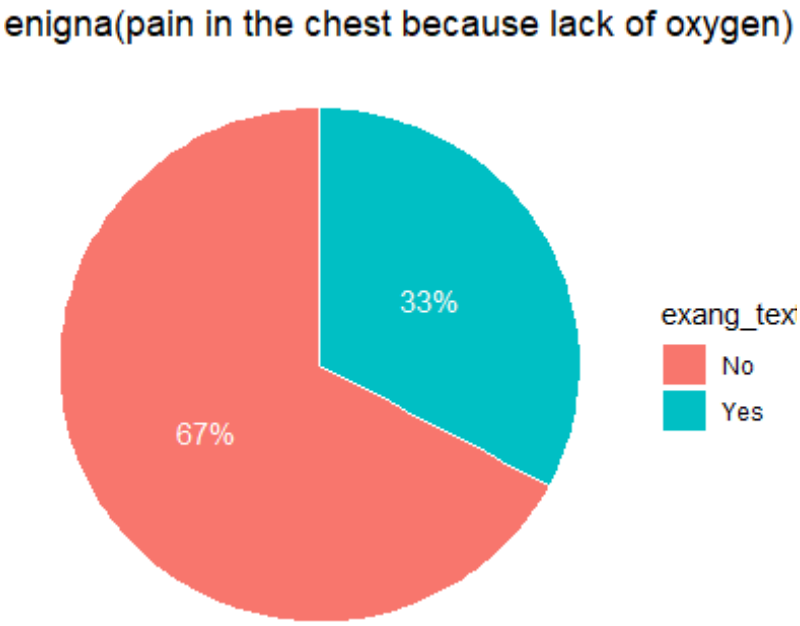

```
> cat("standard deviation",sd(heart_disease$thalach))
standard deviation 22.90353
```

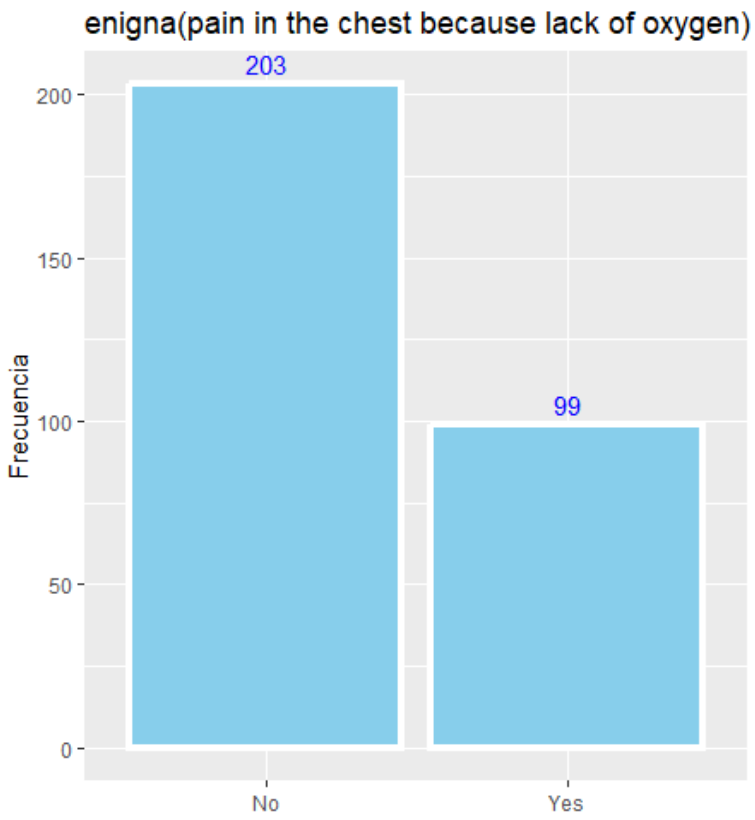
La maxima frecuencia cardiaca que obtuvo una persona cuando estaba haciendo el test de ejercicio fue 202 y el minimo que logro obtener una persona fue 88 en si el 50 % de las personas obtuvieron mas de 152.5 de frecuencia cardiaca y el otro 50 % obtuvo menos de 152.5 de frecuencia cardiaca, la media fue de 149 como el 75 % de los pacientes tuvieron una frecuencia de 133-202 la variacion en cuanto a la media de frecuencia cardiaca fue 23± osea los valores de frecuencia cardiaca en cuanto a la media varian entre el intervalo de 126 y 172.



El intervalo con mas frecuencia fue de 140 – 160 heart rate con un total de 154 paciente ,en si la frecuencia de personas va subiendo desde el heart rate 88 hasta el heart rate 150 pero apartir de ahí empieza a bajar la frecuencia de personas,y se desploma apartir del heart rate 180 hasta llegar a 200 heart rate con1 persona

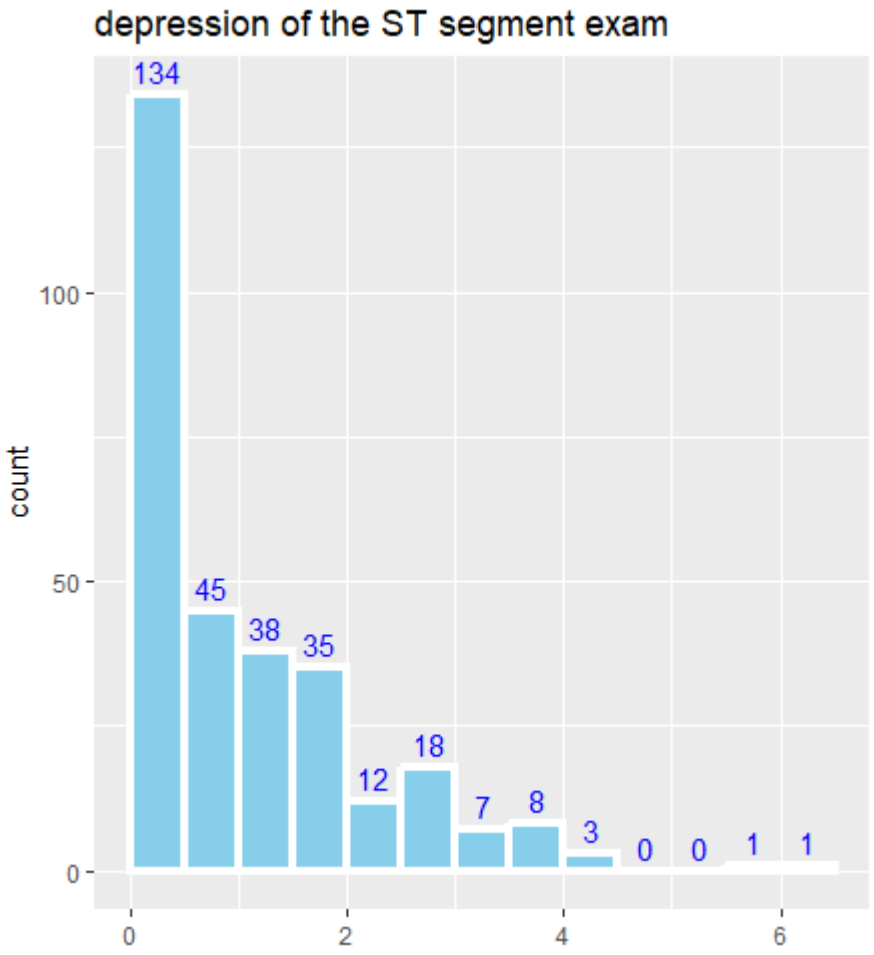
from working out the person had an enigna?





67 % de los pacientes no sufrieron de dolor de pecho por falta de oxigeno mientras el otro restante 33 % si sufrio de anigna, en frecuencia esto seria no = 203 , si = 99 pacientes

Depression of the st segment exam



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.000	0.000	0.800	1.043	1.600	6.200

Depression of the st segment exam **= 0.0 → no depression = normal**

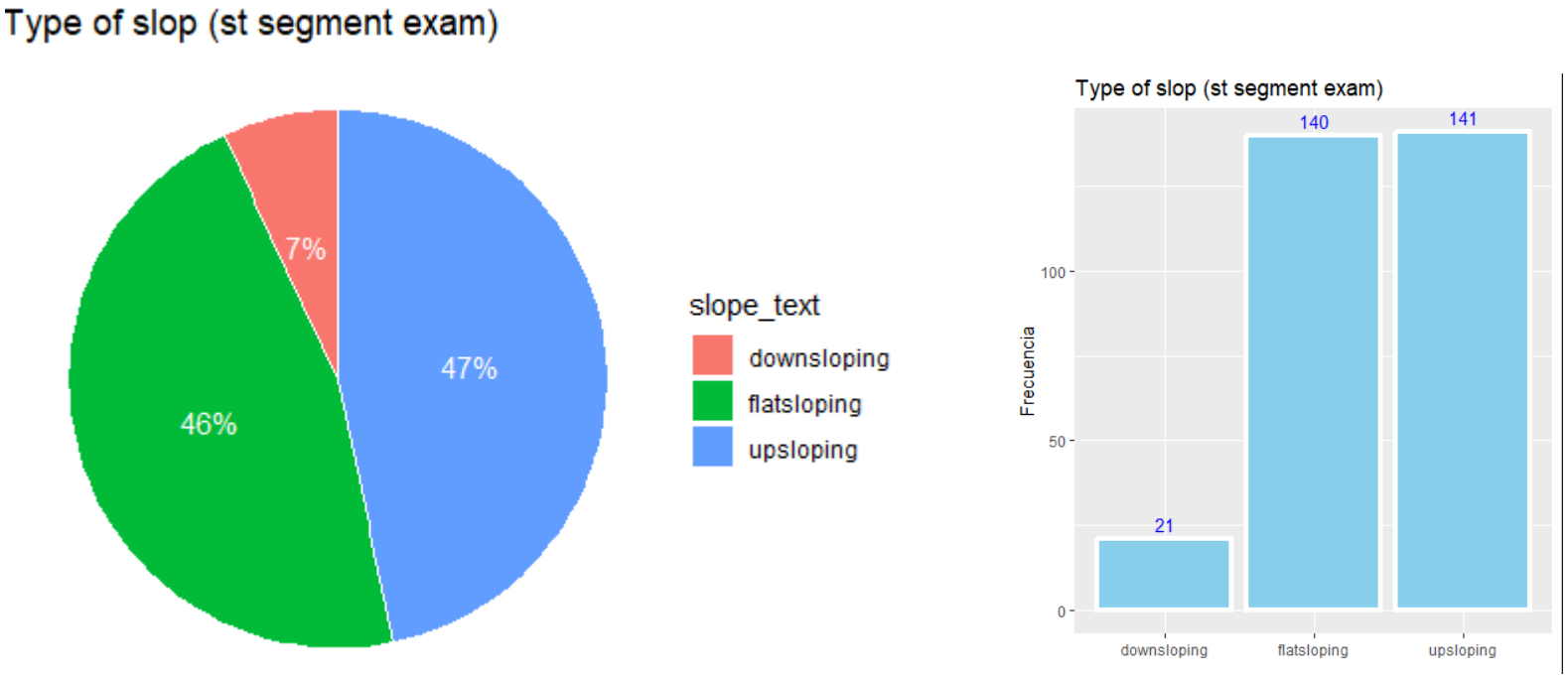
Depression of the st segment exam **= 1.5 → moderate depression**

Depression of the st segment exam **= 3.0 → high depression = higher risk**

La mayoria de pacientes (134) en este test tuvieron 0 de medicion y según la escala del st segment seria una depresion normal -- no riego

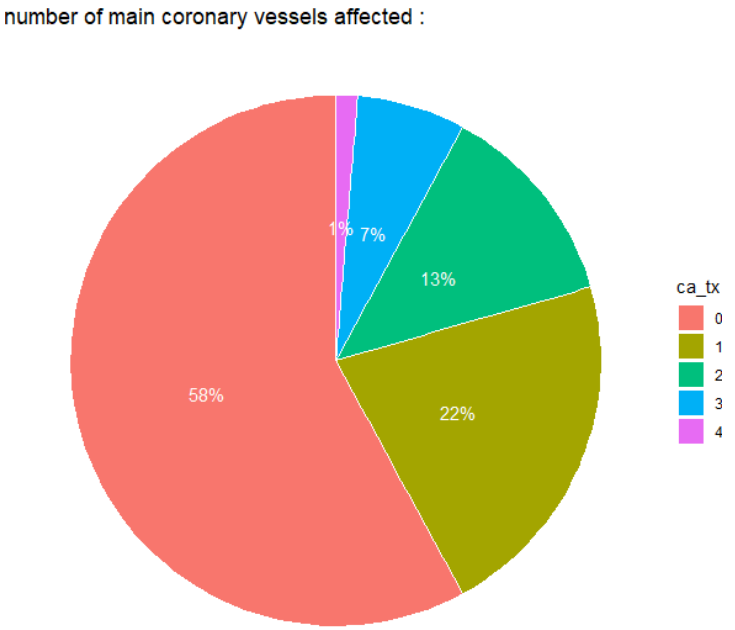
En total serian 217 persona que no estan en riesgo desde el intervalo [0,1.0], el numero de personas que sufren de una depresion moderada— riesgo moderado son 65 intervalo estos pacientes estan en el intervalo [1.5,2.5], los pacientes que estan un riesgo elevado y sufren de una depresion en el corazon tambien elevada son que tienen una medicion de 3.0 >=, el numero de estos pacientes son 20 , en si la media de medicion del examen st depression es de 1.043 osea los pacientes en promedio no sufren de una depresion de el st segment exam el 75 % de los pacientes estan entre una medida de 0 – 1.6 y el maxima medicion registrada a un paciente fue de 6.2 como tambien la medicion con respecto a la media varia en 1.16 ±

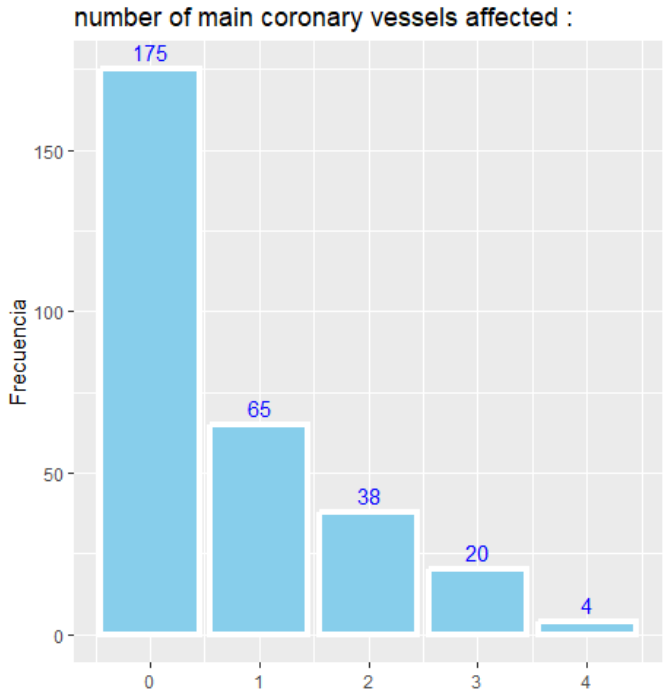
Type of slope the patient had during the st segment exam:



El downsloping es la condicion mas grave el 7 % o una frecuencia de 21 pacientes obtuvieron este resultado, ya en su parte media de seriedad y riesgo estuvo el flatsloping el cual el 46 % de los pacientes tuvieron este diagnostico o en total 140 pacientes y el que no representa ningun riesgo el upsloping tuvo un 47 % o 141 pacientes

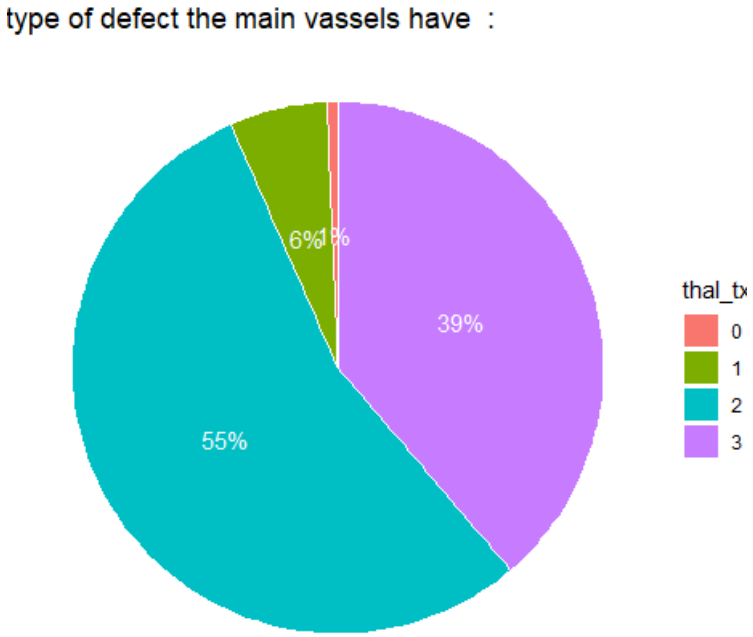
number of main coronary vessels affected :

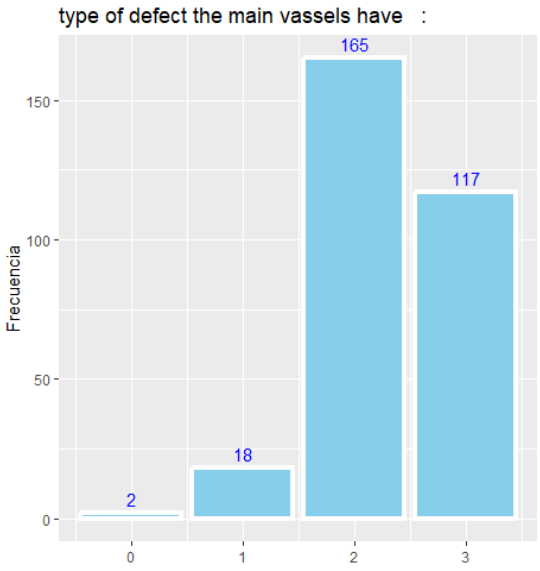




El porcentage de pacientes que no presentaron ningun vaso sanguineo afectado en el examen son 58 % o 175 pacientes los que presentaron un vaso sanguineo afectado fueron el 22 % de los pacientes o 65 pacientes de 2-3 vasos sanguineos afectados (grave) fue un total de 20 % o 58 pacientes y ya por ultimo los que presentaron todos los vasos coronarios princiaples afectados fueron 4 personas o el 1 % de los pacientes (muy grave)

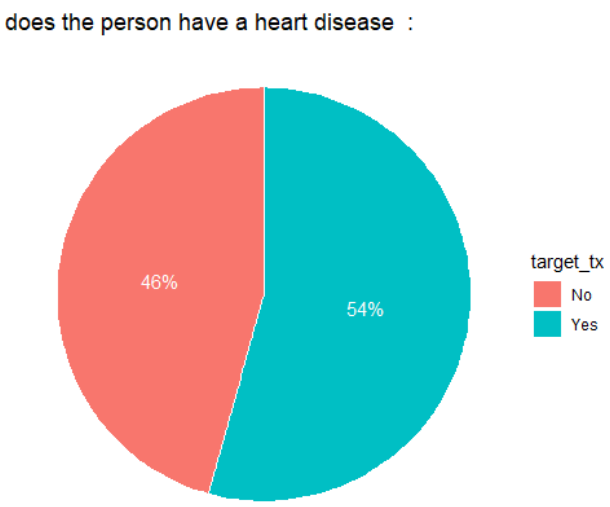
type of defect the main vassels have :





En los valores 0-1 signfica que hay una condicion normal el 7 % de los pacientes obtuvieron este resultado o en frecuencia seria 20 los pacientes que presentaron un valor tipo 2 significa que tienen un defecto NO REVERSIBLE osea fijo en los vasos principales en total fueron el 55 % de los pacientes o en una frecuencia de 165 y el valor 3 significa que hay un defecto en los vasos coronarios pero es revertible en total fueron el 39 % de los pacientes o una frecuencia de 117 osea ya en conclucion pacientes que presentaron algun defecto sea reversible o no fueron el 94 % de los pacientes la moda fue tener un defecto que no es reversible

Does the person have a heart disease ?



El 46 % de los pacientes al final del examen no se les diagnostico una enfermedad del corazon y al otro 54 % si se le diagnositco una enfermedad del corazon en frecuencia seria 138=no 164= si

UNIVARIATE ANALYSIS CONCLUSIONS

Which variables had outliers? Could those variables affect the post-analysis?:

Las variables que presentaron outliers:

1. resting blood pressure(trestbps),
2. cholestherol (chol) ,
3. maximun heart reat achieved(thalach)

resting blood pressure(trestbps):

Empezando con la presión arterial en reposo, se consideró como outlier cualquier valor superior a 170. Sin embargo, no se observa que estos valores afecten significativamente el análisis, ya que esta es una variable de medición y la variabilidad en la presión arterial es natural en el contexto de las enfermedades cardíacas. Además, los valores no se alejan demasiado de los que no se consideran outliers. La variabilidad en esta variable, especialmente en los niveles elevados de presión arterial, es importante, ya que contribuye a definir si una persona tiene o no una enfermedad cardíaca. Por lo tanto, se decidió mantener los valores elevados en el análisis

Cholestherol (chol)

Aquí hay un outlier que destaca por estar muy aparte del valor maximo que no se consideraria outlier que fue 360 este valor outlier fue de 564 mirando a detalle se trata del registro de una mujer de 67 años viendo sus otras medidas para haber si habia una incogruencia vemos que efectivamente la mujer si sufre de un heart disease target=1 el colesterol al ser una medida con mucha variabilidad según el sobrepeso de una persona (puede variar mucho hacia valores altos) y las otras medidas de la mujer tambien son moderadas,riesgosas conclui que iba a dejar este valor por ser muy informativo en este estudio de heart disease

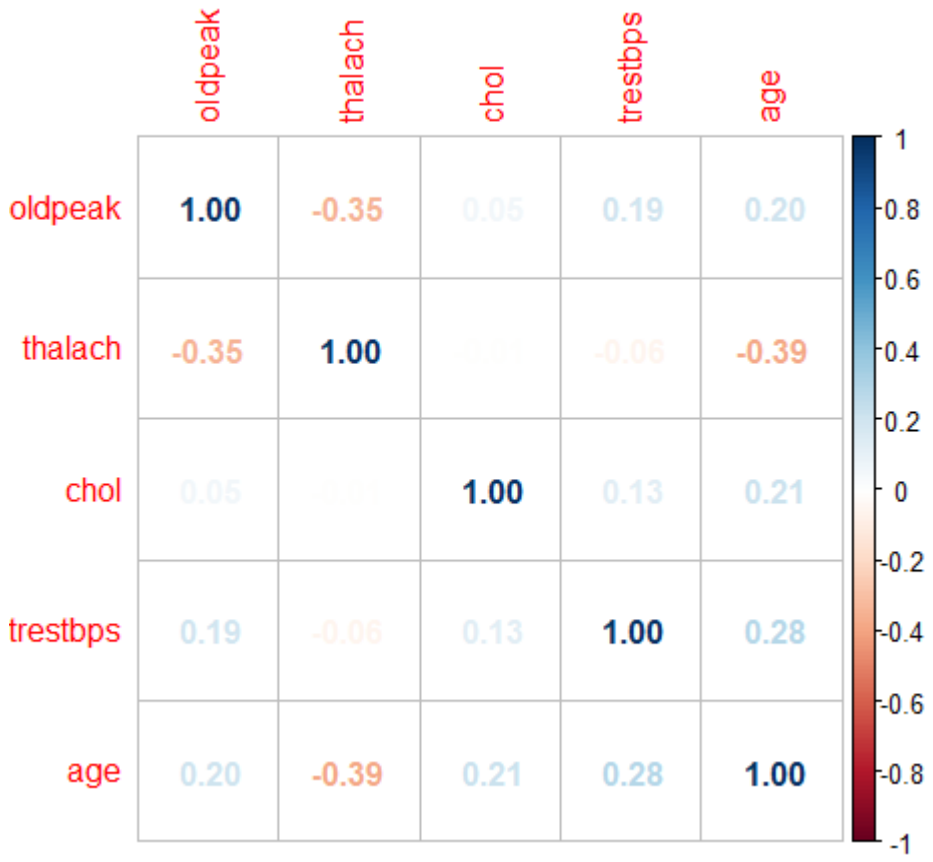
maximun heart rate achieved(thalach)

el outlier fue de 71 de medicion despues de una prueba fisica aquí si hay muchas incongruencias para empezar la persona en efecto no presento heart disease pero tiene 67 años y si miramos registros de esta medicion ” Incluso atletas de élite con gran capacidad aeróbica rara vez bajan de 120-130 lpm en esfuerzo máximo” no sabemos que tanto esfuerzo puso esa persona a la hora del test pero es un valor muy imposible-anormal Un valor de 71 equivaldría a una frecuencia en reposo (no en ejercicio), lo que contradice el propósito de una prueba de esfuerzo entonces hay 2 opciones en efecto la persona no hizo el test bien lo que llevaria a que no es un registro confiable para nuestro analisis posterior (machine learning) o el dato fue incorrectamente introducido por la tanto he decidio eliminar este registro.

Was there a variable that showed unusual behavior?

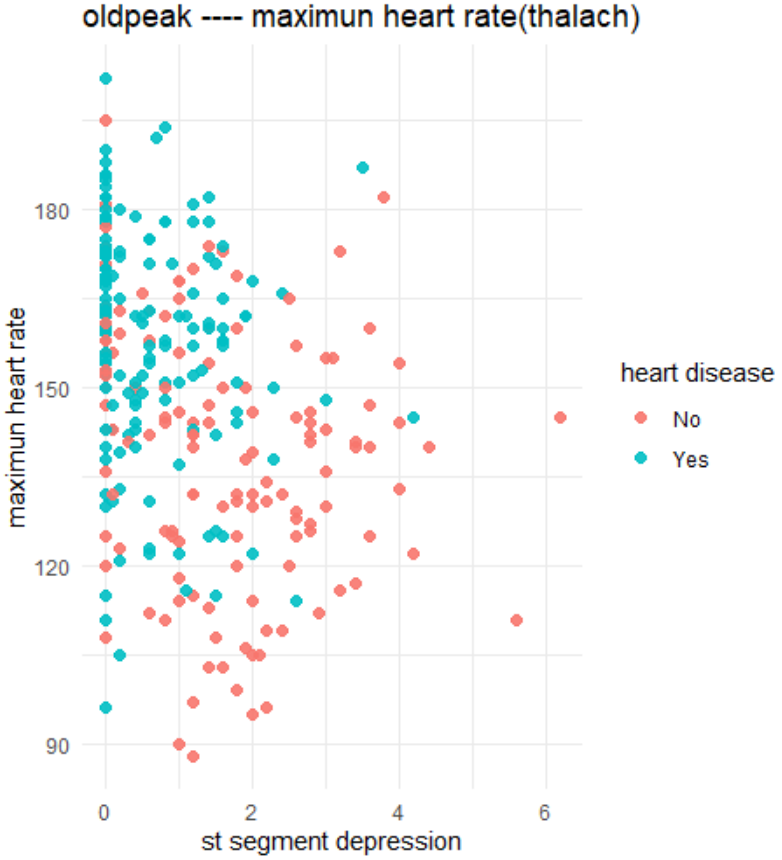
No, en todas las variables no vi algo que se fuera de contexto con lo que se busca y es entender si hay un heart disease o no,aparte de los outliers no vi asi marcados errores de medicion o cosas sin contexto o sentido con lo que se busca

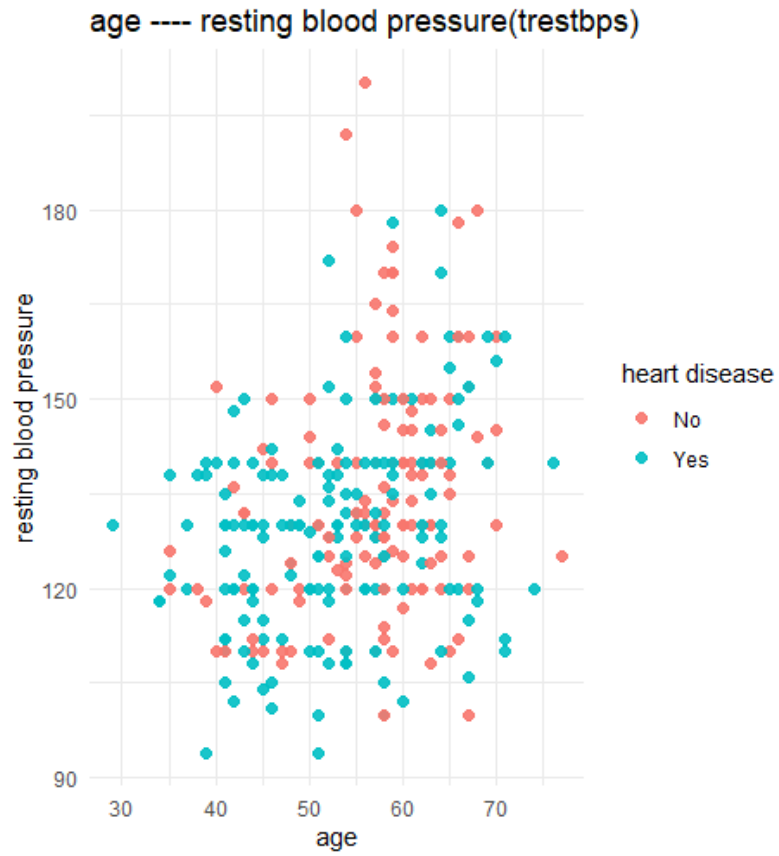
BI-MULTI VARIATE ANALYSIS



Entre variables numericas continuas no hay una relacion la que mas relacion mostro pero no fue significativo fue la variable age con thalach hay una relacion inversa entre ellas si sube una baja la otra osea si el maximum heart rate sube la edad baja y asi alrevez pero no llega a ser una relacion prominente o que de verdad exista


age ---- maximun heart rate(thalach)
maximun heart rate
180
150
120
90
30
40
50
60
70
age
heart disease
No
Yes


oldpeak ---- maximun heart rate(thalach)
maximun heart rate
180
150
120
90
0
2
4
6
st segment depression
heart disease
No
Yes


age ---- resting blood pressure(trestbps)
resting blood pressure
180
150
120
90
30
40
50
60
70
age
heart disease
No
Yes


cholestherol ---- age
cholesterol
500
400
300
200
30
40
50
60
70
age
heart disease
No
Yes

Mirando las variables que tuvieron mejor relacion en el analisis de correlacion vemos que por ejemplo que en age-maximun hear rate vemos que efectivamente entre mayor sea la frecuencia cardiaca con la que acaba el paciente mas probabilidad de heart disease pero en si la edad no tiene nada que ver con ello no hay relacion entre esas dos variables,

En el examen de st segment depression con maximun heart rate vemos que no hay ninguna relacion lo que marca si hay un heart disease es el maximum heart rate no el st segment depression en esta relacion

En age con resting blood pressure vemos que no hay nada de relacion con respecto al heart disease amabas variables van por su cuenta y lo mismo con chol y age no se ve tampoco una relacion

¿Qué variables no parecen aportar informacion entre las variables continuas --- enfocado en el target ?

Veo que age en ningun contexto tuvo relevancia pero con maximun heart rate veo que si hay una fuerte relacion con el heart disease cholestherol y resting blood pressure tocaria verlas mas adelante a profundidad para ver si hay relacion con el heart disease

CHI- SQUARE

Variable	P_Valor	Relación_Significativa
sex	1.871595e-06	Sí
cp	2.864962e-17	Sí
fbs	7.410028e-01	No
restecg	6.252540e-03	Sí
exang	6.415093e-14	Sí
slope	9.175160e-11	Sí
ca	2.215886e-15	Sí
thal	1.621748e-18	Sí

aqui se hizo un test chi -cuadrado para mirar la relacion entre variables categoricas en especial de cada variable categorica la relacion que tendria con el target.

Es significativa la relacion si p_valu es < nivel de significancia 0.05 entonces en efecto hay una relacion significativa si es al contrario no hay ninguna relacion.

La unica variable que no mostro relacion con target fue fbs “fating blood sugar>120”

ANALYSING EACH CATEGORICAL VARIABLE VS NUMERCIAL

Age Vs Categorical :

Vemos primero que la edad con el sexo se ve muy influyente a la hora de decir si alguien tiene heart disease pordemos ver que las mujeres si las agrupamos por edad casi tdas tienen un heart disease al contrario del hombre pero la edad no se ve clara aquí en la mujer por que no importa la edad todas tienen heart disease y en hombres a mayor edad menor heart disease

VS Chest pain(cp):

Los que sufren del tipo de dolor de pecho TP su mayoria no tiene un heart disease pero en los otros dolores de pecho como el (ATP ,NO) prevalece un diagnostico de heart disease positivo según la edad ,en el ultimo que dice que no hay dolor de pecho

asn(asintomatico) tambien por edad prevalece un diagnostico positivo no se ve un patron claro en la edad a la hora de la calsificacion

VS Fasting blood sugar > 120 (fbs)

Aquí en cuanto la edad y el nivel de azucar en ayunas no se ve un clara relacion que distinga un heart disease positivo

VS Electrocardiographic results (restecg)

Se puede ver que hay un poco mas de pacientes con heart disease cuando el resultado fue leve (gravedad) en comparacion con normal (no hay gravedad) en su diagnostico grave =2 como no hay suficientes datos no se puede inferir del todo pero mostro que a mas edad hay presencia de un heart disease .

VS From working out the person had an enigna (exang)

Estas dos variables no tiene una relacion clara por que cuando la persona sufrio de dolor de pecho despues de hacer ejercicio eso comparado con la edad dio que casi toda esa gente no tuvo un heart disease sin importar la edad, pero en el caso contrario cuando la persona no tuvo un dolor de pecho despues de hacer ejercicio hubo muchos que tuvieron un heart disease sin importar la edad .

VS Type of slop (slop)

En el type of slop 2 que es un caso normal upslopping esta concentrado la mayoria de gente con heart disease ya entre mas grave sea el caso menos gente hay con heart disease osea hay una relacion inversa tiene el mismo comportamiento de la variable from working out the person had an enigna

VS Number of coronary vassels affected (ca)

aqui tambien la relacion con respecto a la edad es inversa al igual que la variable (exang,slop) entonces se podria decir que con respecto a la edad si un paciente cada vez tiene mas vasos coronarios afectados no habria relacion alguna en cuanto a un heart disease positivo

VS Type of defect of the main vassels (thal)

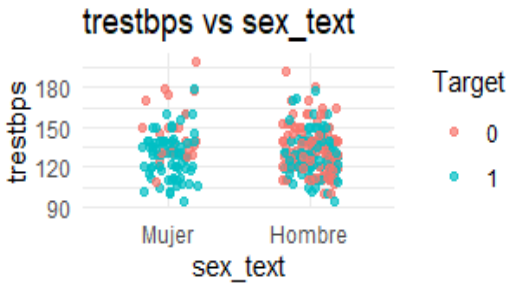
Esta variable con respecto a la edad si tiene una relacion clara a la hora de diagnosticar un heart disease por que su valor 2 que seria un defecto no arreglable (muy grave) esta todos los pacientes con heart disease cuando el efecto es reversible valor 3 casi no hay pacientes con heart disease

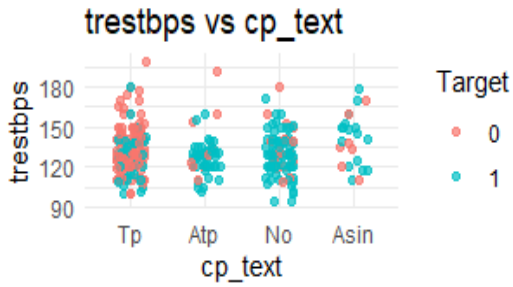
CONCLUCIONES FINALES AGE :

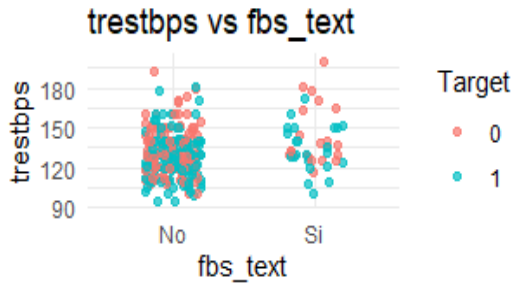
Despues del analisis hecho a esta variable no diria yo que influye mucho a la hora de un heart disease, cuando se relaciona con variables numericas continuas no hay relacion alguna y ya con categoricas vemos que aunque hay variables que tiene mucha relacion como el genero el tipo de defecto de los vasos principales, un poco de relacion en el chest pain y en elctrocardiograma en las otras variables no tiene efecto alguno mas bien hace un efecto a la inversa

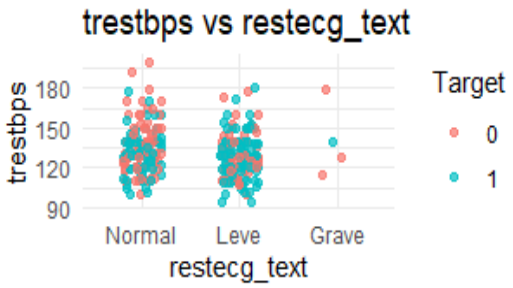
En 8 variables en total no tendria relacion alguna a predecir un heart disease en 2 variables tendria una fuerte relacion y en las otras 2 variables tendria una relacion leve -media

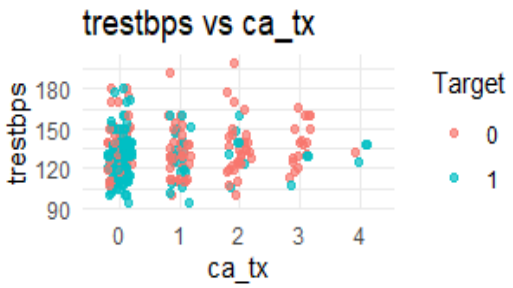
Resting blood pressure VS Categorical :

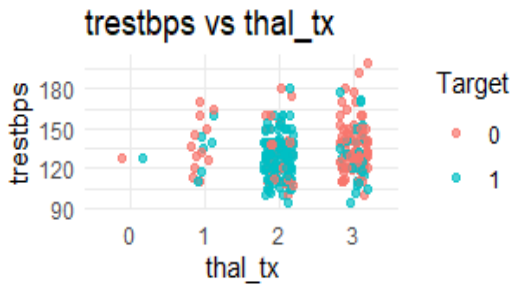












VS Sex:

Hay una muy buena relacion con el genero para predecir un heart disease siendo la mujer la que mas tiene casos positivos y el hombre casos negativos

VS Chest pain(cp):

Tambien se podria decir que hay una buena relacion en siertos dolores de pecho

VS fasting blood sugar(fbs) :

No hay una relacion clara entre estas dos variables al igual que la edad esta variable de fasting blood sugar se ve que en si no tiene relacion con ninguna variable y menos con un heart disease

VS Electrocardiographic results (restecg)

Aqui no veo una relacion clara se ve que cuando el diagnosito fue leve (gravedad) en comparacion con normal (no hay gravedad) hay un poco mas de pacientes con heart disease pero no se ve un efecto claro y el caso mas grave no tiene casi datos como para poder inferir

VS From working out the person had an enigna (exang)

Aqui hay una relacion inversa cuando los pacientes no tienen un enigna (dolor de pecho despues de hacer ejercicio) esta la mayoria de personas con heart disease cuando se agrupa con el resting blood pressure.

VS VS Type of slop (slop)

Pasa lo mismo que con exang una relacion inversa en la que no se puede relacionar una cosa con la otra para inferir un heart el caso normal de slope upslopping tiene la mayoria de personas con heart disease

VS Number of coronary vassels affected (ca)

Lo mismo que (slop,exang) relacion inversa a mas vasos coronarios afectados menos persona con heart disease cuando se compara con resting blood pressure

VS Type of defect of the main vassels (thal)

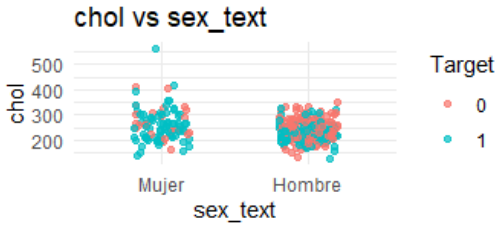
Esta variable con respecto al resting blood pressure si tiene una relacion clara a la hora de diagnosticar un heart disease por que su valor 2 que seria un defecto no

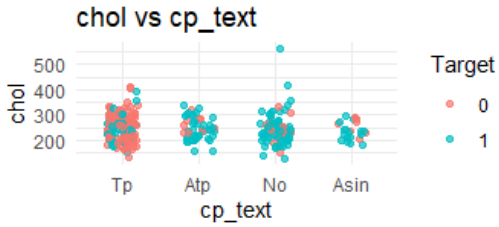
arreglable (muy grave) esta todos los pacientes con heart disease cuando el efecto es reversible valor 3 casi no hay pacientes con heart disease

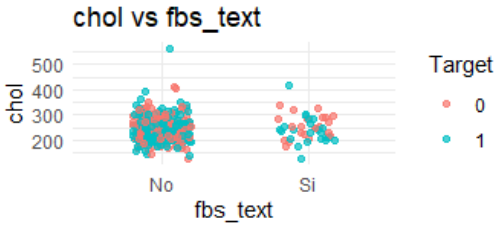
CONCLUSIONES RESTING BLOOD PRESSURE TRESTBPS:

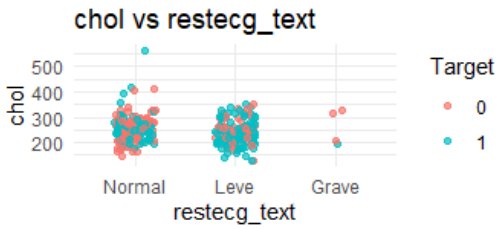
Maneja una relacion fuerte con 3 variables(thal,sex,cp) en total tambien esta variable tiene relaciones inversas con 3 variables(ca,slop,exang) en total, y por ultimo tiene una relacion nula con (fbs) y una relacion leve con (resttecg)

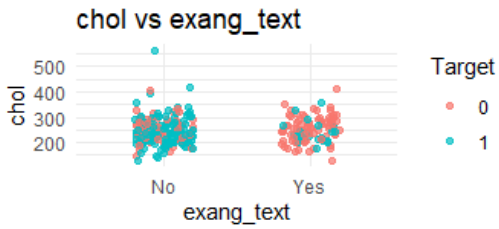
Cholesterol VS Categorical :

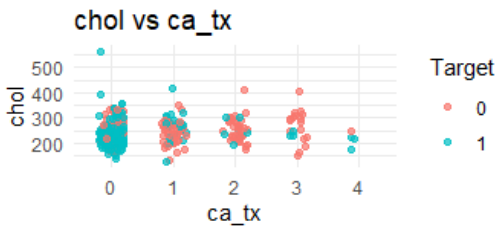


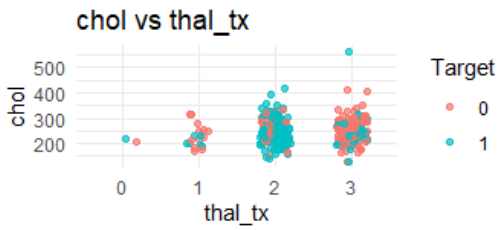












VS Sex :

Hay una muy buena relacion con el genero para predecir un heart disease siendo la mujer la que mas tiene casos positivos y el hombre casos negativos

Vs cp

Los que sufren del tipo de dolor de pecho TP su mayoria no tiene un heart disease pero en los otros dolores de pecho como el (ATP ,NO) prevalece un diagnostico de heart disease positivo según el cholestherol ,en el ultimo que dice que no hay dolor de pecho asn(asintomatico) tambien por cholestherol prevalece un diagnostico positivo

VS fasting blood sugar(fbs) :

No hay una relacion clara entre estas dos variables al igual que el resting blood pressure esta variable de fasting blood sugar se ve que en si no tiene relacion con ninguna variable y menos con un heart disease

VS Electrocardiographic results (restecg)

Aqui veo una relacion medio clara se ve que cuando el diagnosito fue leve (gravedad) en comparacion con normal (no hay gravedad) hay un poco mas de pacientes con heart disease y tambien a valores altos hay heart disease el caso mas grave no tiene casi datos como para poder inferir

VS From working out the person had an enigna (exang)

Aqui hay una relacion inversa cuando los pacientes no tienen un enigna (dolor de pecho despues de hacer ejercicio) esta la mayoria de personas con heart disease cuando se agrupa con el chol

VS Type of slop (slop)

Pasa lo mismo que con exang una relacion inversa el caso normal de slope "upslopping" tiene la mayoria de personas con heart disease siguiendo el mismo patron visto en esta variable esta etiqueta siempre tiene la mayoria de personas con heart disease

VS Number of coronary vassels affected (ca)

Lo mismo que (slop,exang) relacion inversa a mas vasos coronarios afectados menos persona con heart disease cuando se compara con cholestherol

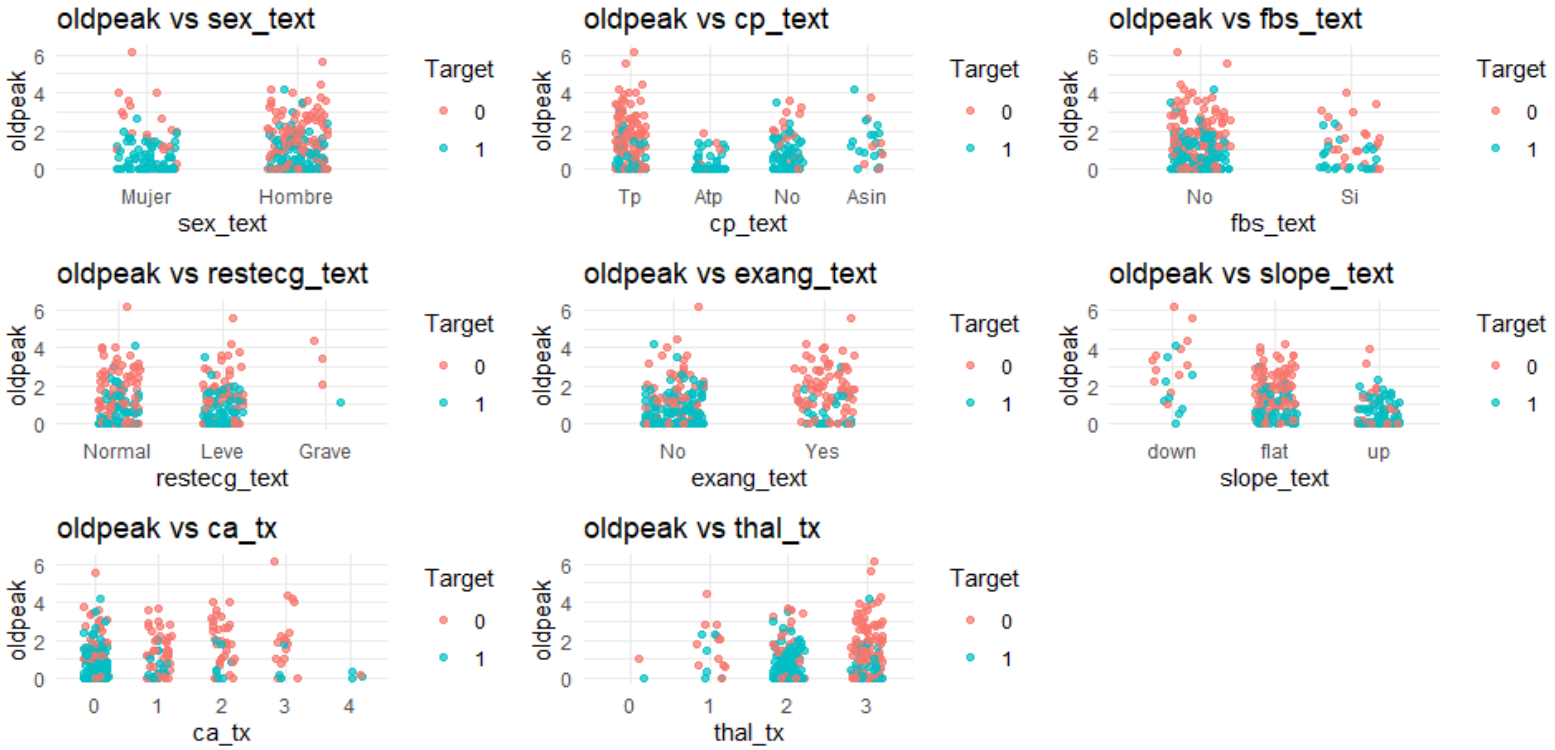
VS Type of defect of the main vassels (thal)

Esta variable con respecto al chol si tiene una relacion clara a la hora de diagnosticar un heart disease por que su valor 2 que seria un defecto no arreglable (muy grave) esta todos los pacientes con heart disease cuando el efecto es reversible valor 3 casi no hay pacientes con heart disease

CONCLUCIONES CHOLESTHEROL

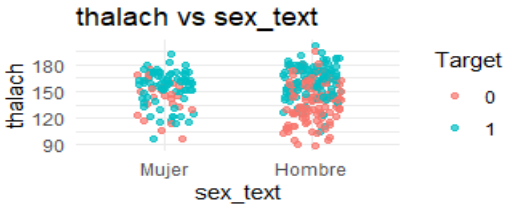
Se comporta parecido que resting blood pressure "trestbps" cuando se compara con variables categorica son variables que estan correlacionadas pero esta variable cholestherol presenta matices de mejor clasificacion que trestbps haciendo congruencia que a mas cholestherol mas presencia de un heart disease a valores altos de cholestherol hay mas heart disease .

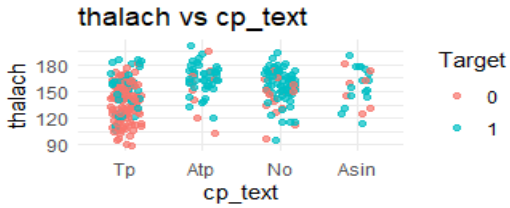
ST Segment depression test vs Categorical

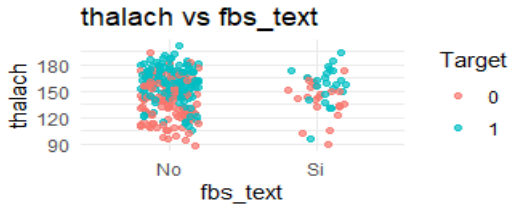


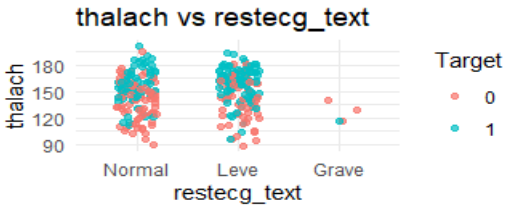
Esta variable sigue la misma logica de chol , resting blood pressure pero con la diferencia de que los casos positivos de heart disease y los no positivos no se clasifican mejor que en la variable choletherol pero si se clasifica mejor que trestbps cuando se comparo con la variable maximum heart rate tuvo una de las relaciones mas fuertes entre las variables continuas pero no tan fuerte para ser significativa a valores bajos se ve que en variables como sex o cp o thal,restecg ,fbs le va sobresaliente a al hora de clasificar los casos positivos en si esta variable a valores bajos clasifica mucho la etiqueta heart disease positiva

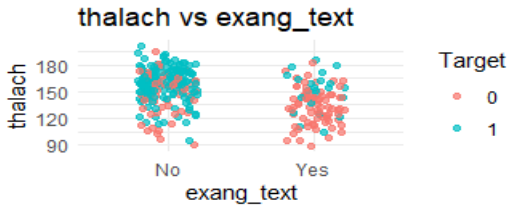
Maximun heart rate VS Categorical

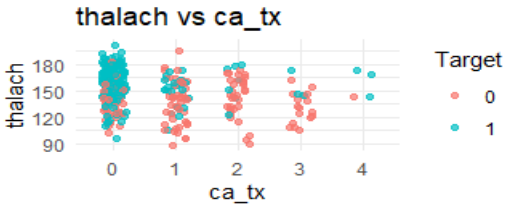


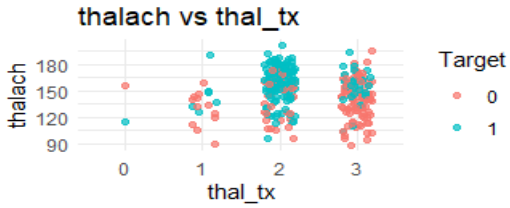












entre variables numericas fue la que mejor resultados tuvo mostrando una relacion con cada variable numerica a la hora de mirar un heart disease positivo muy bueno,con las variables categoricas vemos que en sex tiene su relacion como tambien en chest pain ,thal llega a tener una relacion tambien buena y congruente de resto pues relaciones inversas o leves que ciertas variables categoricas han ido mostrando en el analisis se destaca por que en los valores altos siempre presenta un heart disease lo cual hace congruencia con esta variable

si es una variable que dejaria por su buen rendimiento a la hora de clasificar un heart disease positivo entre variables numericas y su congruencia en valores altos en las varibales categoricas

Regresion logistica coeficientes:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.883623   2.631474   1.476 0.139988
age         -0.004914   0.023571  -0.208 0.834851
sex         -1.713658   0.471388  -3.635 0.000278 ***
cp           0.842510   0.186846   4.509 6.51e-06 ***
trestbps    -0.021062   0.010422  -2.021 0.043285 *
chol        -0.003935   0.003834  -1.026 0.304732
fbs          0.067690   0.530264   0.128 0.898424
restecg      0.410787   0.350983   1.170 0.241844
thalach      0.020848   0.010783   1.933 0.053190 .
exang       -1.002726   0.412926  -2.428 0.015168 *
oldpeak     -0.550055   0.215779  -2.549 0.010798 *
slope        0.576302   0.350952   1.642 0.100567
ca          -0.808063   0.201842  -4.003 6.24e-05 ***
thal        -0.894304   0.290616  -3.077 0.002089 **
```

Aquí por medio de regresion logistica se miro que tan estadisticamente influyente seria cada variable a la hora de explicar la variabilidad de la variable dependiente “target” vamos a manejar un nivel de significancias de 0.05 y según lo que de el p-value de cada variable (Pr(>|z|) miramos que tan significativo es esa variable con respecto al target “Si (Pr(>|z|) < 0.05, se podria decir que esa variable tiene un efecto estadísticamente significativo sobre la target.” (se uso

regresion logistica para hacer esto por que el target es binario 0,1 “calculo de probabilidad”)

Variables significativas ((Pr(>|z|)< 0.05)

sex, **cp**, **trestbps**, **exang**, **oldpeak**, **ca**, **thal, thalach,chol**

Variables NO significativas ((Pr(>|z|)>0.05)

age, **fbs**, **restecg**, **slope**

Analisis de colinealidad :

Aquí se miro si dos o más variables independientes (predictoras) están altamente correlacionadas entre ellas, ya que si tenemos este tipo de variables sería muy difícil para el modelo inferir de forma precisa cuál variable es la que realmente está generando el efecto sobre el target**.**

Se uso la librería VIF (Variance Inflation Factor)**.**
Un VIF mayor a 5 o 10 nos indica que hay colinealidad alta-grave, y que deberíamos considerar eliminar o transforma**r** esa variable.

Resultados:

```
      age       sex        cp  trestbps      chol       fbs   restecg  thalach     exang   oldpeak     slope
1.388012  1.355981  1.280475  1.136683  1.245273  1.135565  1.077173  1.430583  1.159000  1.421626  1.485145
      ca      thal
1.134948  1.077146
```

Aquí podemos ver que según este test no hay ninguna correlacion o colinealidad “fuerte ”entre variables al estar en un rango sano de 1-1.43

CONCLUSIONES:

Age:

Al analizar la variable age en relación con otras variables numéricas y categóricas, no se evidenció una asociación clara con la presencia de enfermedad cardíaca. Es decir, no se observó que el hecho de que una persona sea mayor o no tenga un impacto relevante en la aparición del heart disease. Además, en la prueba con coeficientes de regresión logística, esta variable también fue descartada por no aportar significativamente al modelo. Por lo tanto, no considero aprobarla.

Sex:

Esta variable si presento relacion según sea el genero a la hora de predecir un heart disease yo la dejo por que su relacion siempre fue muy buena en el chi square mostor que es relevante para el target

Chest pain (cp):

Esta variable siempre mostro una muy buena variabilidad según el tipo de dolor de pecho a la hora de clasificar un heart disease o no y en la prueba de chi square mostro que era relevante para el target

Resting blood pressure (trestbps), Chol , st segment depression (Oldpeak):

Estas variables tenia resultados muy parecidos exceptuando un poco trestbps que no deja muy en claro la clasificacion de heart disease como en las otras variables por ejemplo cholestherol mantiene una relacion lineal de que a mas cholestherol mas heart disease en st segment depression los valores bajos los

maneja sobresalientemente esta variable de trestbps maneja un comportamiento parecido a chol entonces decidi dejar chol por rendir mejor en la clasificacion de casos positivos como tambien deicidi dejar oldpeak por discernir mejor los valores bajos

Fasting blood (fbs):

Fue una variable que nunca mostro una relacion clara y en el chi -square no paso la prueba

Electrocardiographic results (restecg)

Es una variable que dejo por que hay una clara clasificacion de heart disease si es leve(intensidad de seriedad) o normal (no hay efecto) siempre en leve mostro destacarse a la hora de un heart disease a comparacion de normal aunque en la regresion logistica no paso en el chi square si paso yo si la dejo

Maximun heart rate (thalach)

Esta variable fue la mejor con chol en discernir el target osea que la dejo

From working out the person had an enigna (exang)

Fue una variable que siempre en su valor no hubo muchas personas con heart disease paso las pruebas que se le hicieron de chi-square y regresion logistica entonce yo la dejo

Type of slop (slop)

Es un variable que siempre mostro en upslopping que hay bastantes casos de gente con heart disease en la pruba de chi-square paso y en regresion no pero dicidi dejarla por que siempre en esa etiqueta clasifico muchas personas con heart disease y en su otra etiqueta flatslopping siempre prevalecio el no heart disease

Number of coronary vassels affected (ca)

Tambien la dejo por mostrar variabilidad entre sus etiquetas en cuanto el heart disease paso su test de regresion logistica y chi square

Type of defect of the main vassels (thal)

la variable thal fue muy clara que cuando el valor 2 “no reversable defect” aparece significa que hay un heart disease osea en si esta variable siempre manejo una congruencia que cuando la persona tiene un defecto coronario es por que tiene un heart disease “paso todas sus pruebas chi-square--regresion”

Variables que no aportarian al target : age,trestbps,fbs