Estudiante: David Andrés Valencia García
Profesor: José Daniel Ramírez Soto

1. Utilizar el siguiente set de datos para calcular paso por paso.

x1	x2	x3
4	4	28
2	3	24
2	4	30
3	5	32
1	3	18
3	6	41
3	6	44
0	1	5
1	3	18
0	0	1
5	9	62
1	2	17
2	3	24
1	3	19
3	6	42
4	8	56
4	8	56
3	6	44
5	9	64
1	2	17
1	2	17

1.1. ¿Cuál es la media, mediana y desviación estándar?, y la moda y los valores repeticiones de la moda para los datos categóricos

1.1 Media: $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ } $n=21$

- Para X_1

$\rightarrow \bar{X}_1 = (4+2+2+3+1+3+3+0+1+0+5+1+2+1+3+4+4+3+5+1+1) \cdot \frac{1}{21}$

$\Rightarrow \bar{X}_1 = \frac{49}{21} \Rightarrow \therefore \bar{X}_1 = 2,333$

- Para X_2

$\rightarrow \bar{X}_2 = (4+3+4+5+3+6+6+1+3+0+9+2+3+3+6+8+5+6+9+2+2) \cdot \frac{1}{21}$

$\Rightarrow \bar{X}_2 = \frac{93}{21} \Rightarrow \therefore \bar{X}_2 = 4,428$

- Para X_3

$\rightarrow \bar{X}_3 = (28+24+30+32+18+41+49+5+18+1+62+17+24+19+12+56+44+64+17+17) \cdot \frac{1}{21}$

$\Rightarrow \bar{X}_3 = \frac{659}{21} \Rightarrow \therefore \bar{X}_3 = 31,38$

Mediana: Valor central

• Para X_1:

→ $Me_1 = 0, 0, 1, 1, 1, 1, 1, 1, 2, 2, \textcircled{2}, 3, 3, 3, 3, 3, 4, 4, 4, 5, 5$

$\therefore Me_1 = 2 //$

• Para X_2

→ $Me_2 = 0, 1, 2, 2, 2, 3, 3, 3, 3, 3, \textcircled{4}, 4, 5, 6, 6, 6, 6, 8, 8, 9, 9$

$\therefore Me_2 = 4 //$

• Para X_3

→ $Me_3 = 1, 8, 17, 17, 17, 18, 18, 19, 24, 24, \textcircled{28}, 30, 32, 41, 42, 44, 44, 56, 56, 62, 64$

$\therefore Me_3 = 28 //$

Desviación Estandar $\sigma = \sqrt{\underbrace{\frac{1}{N}\cdot\sum_{i=1}^{N}(x_i-\bar{x})^2}_{Var(x)}}$

- Para X_1

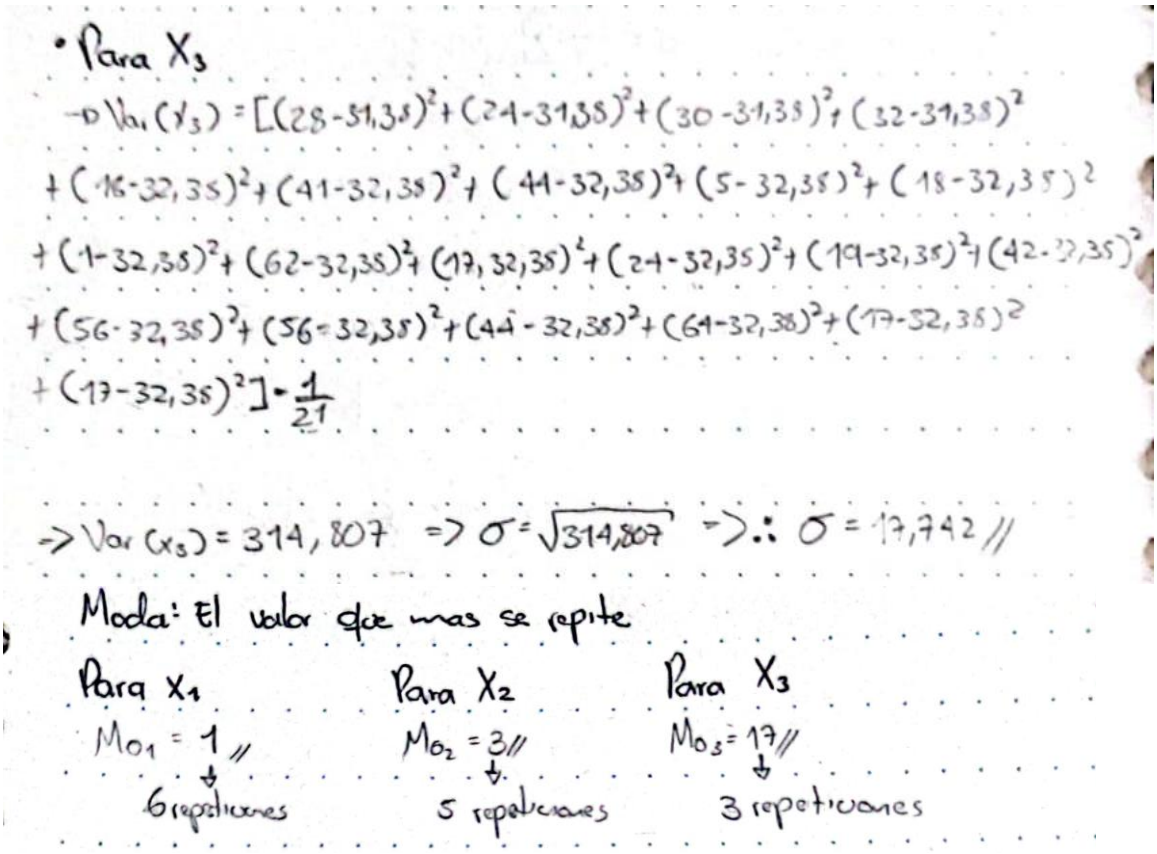
$\rightarrow Var(X_1) = [(4-2,333)^2 + (2-2,333)^2 + (2-2,333)^2 + (3-2,333)^2 + (1-2,333)^2$
$+ (3-2,333)^2 + (3-2,333)^2 + (0-2,333)^2 + (1-2,333)^2 + (0-2,333)^2 + (5-2,333)^2$
$+ (1-2,333)^2 + (2-2,333)^2 + (1-2,333)^2 + (3-2,333)^2 + (4-2,333)^2 + (4-2,333)^2$
$+ (3-2,333)^2 + (5-2,333)^2 + (1-2,333)^2 + (1-2,333)^2] \cdot \frac{1}{21}$

$\Rightarrow Var(X_1) = 2,222 \Rightarrow \sigma = \sqrt{2,22} \Rightarrow \therefore \sigma = 1,49$ //

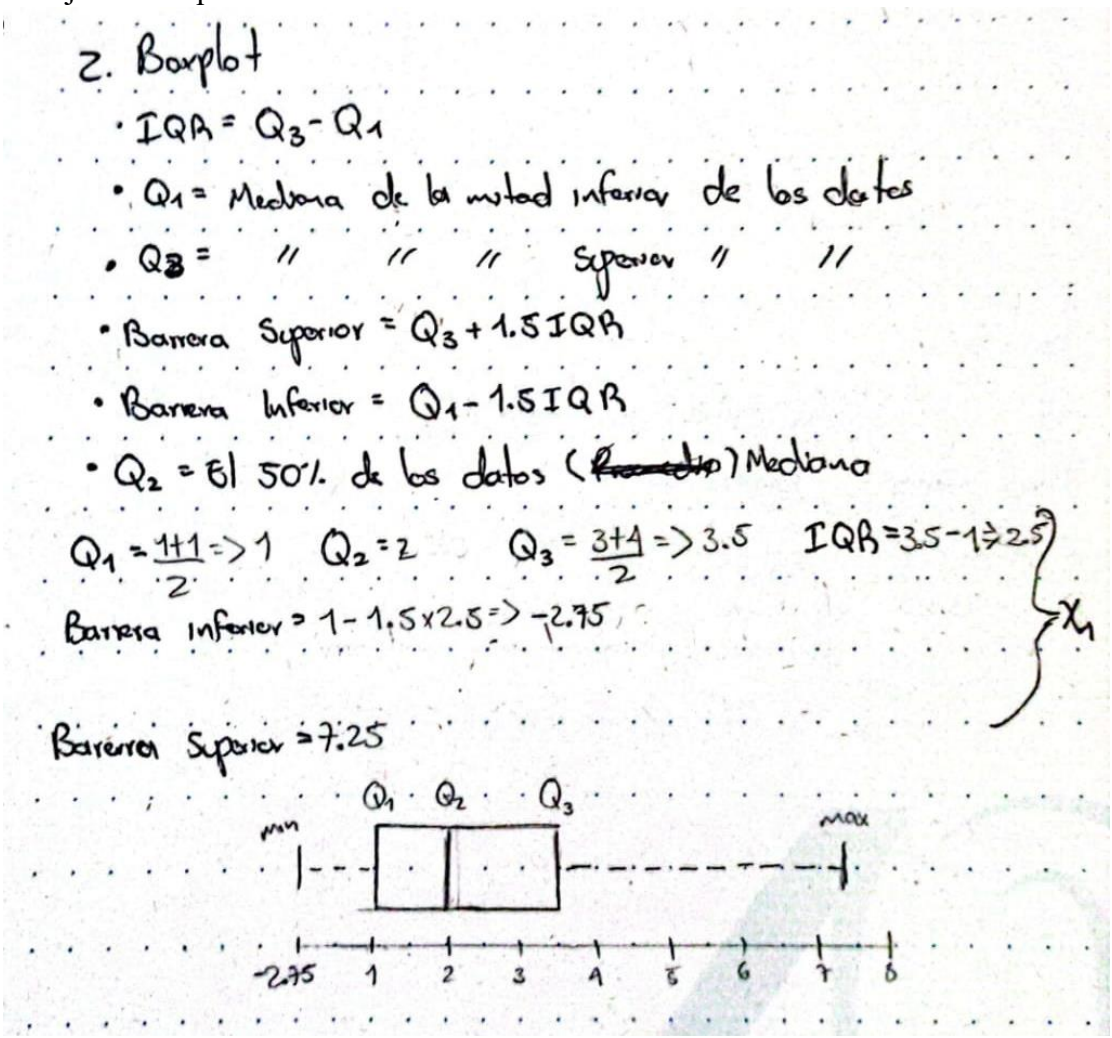
- Para X_2

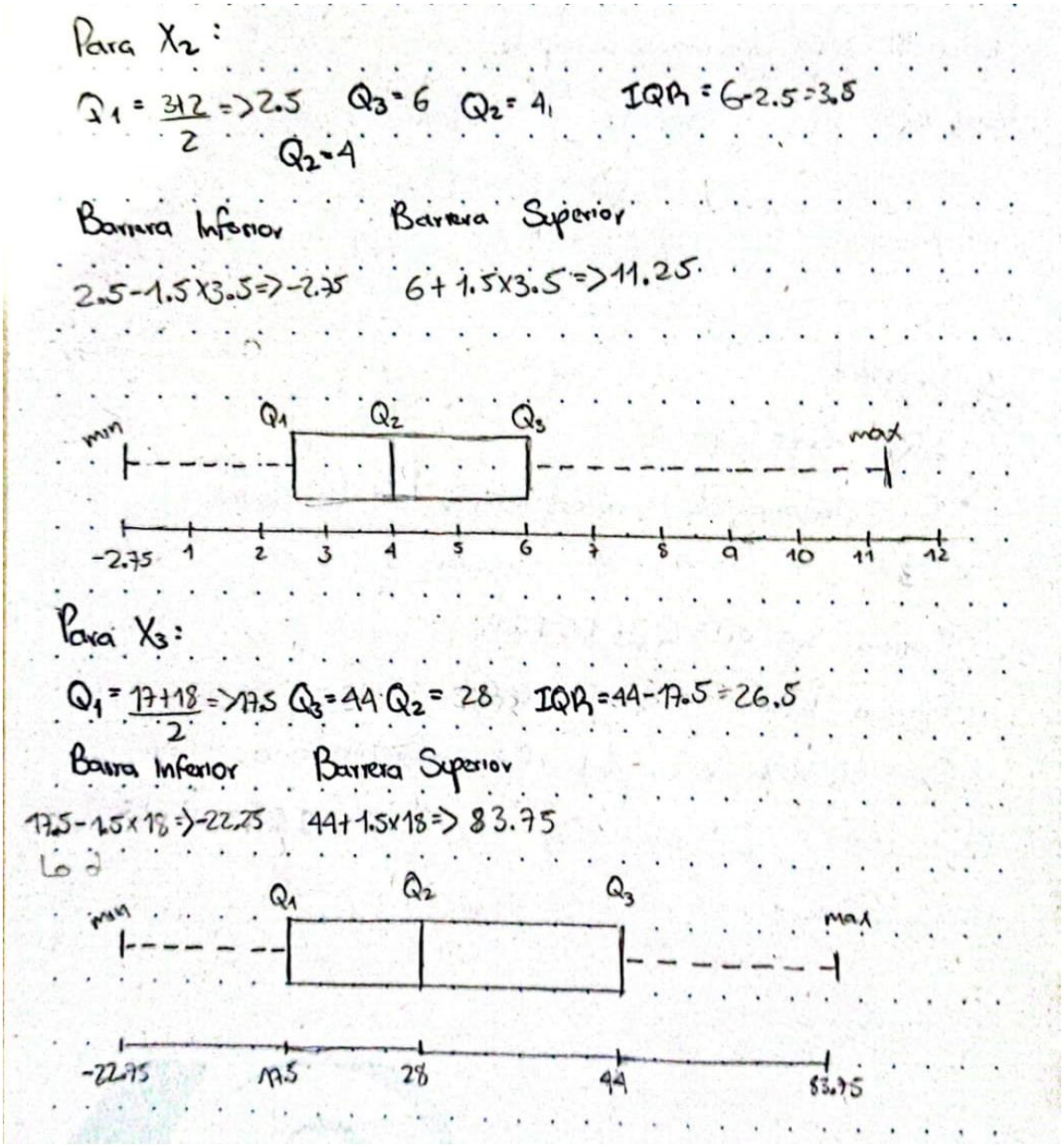
$\rightarrow Var(X_2) = [(4-4,428)^2 + (8-4,428)^2 + (4-4,428)^2 + (5-4,428)^2 + (3-4,428)^2$
$+ (6-4,428)^2 + (6-4,428)^2 + (1-4,428)^2 + (3-4,428)^2 + (0-4,428)^2 + (9-4,428)^2$
$+ (2-4,428)^2 + (8-4,428)^2 + (3-4,428)^2 + (6-4,428)^2 + (8-4,428)^2 + (8-4,428)^2$
$+ (6-4,428)^2 + (9-4,428)^2 + (2-4,428)^2 + (2-4,428)^2] \cdot \frac{1}{21}$

$\Rightarrow Var(X_2) = 6,53 \Rightarrow \sigma = \sqrt{6,53} \Rightarrow \therefore \sigma = 2,555$ //


• Para X_3
$\rightarrow Var(X_3) = [(28-31,38)^2 + (24-31,38)^2 + (30-31,38)^2 + (32-31,38)^2$
$+ (16-32,35)^2 + (41-32,35)^2 + (44-32,35)^2 + (5-32,35)^2 + (18-32,35)^2$
$+ (1-32,35)^2 + (62-32,35)^2 + (17,32,35)^2 + (24-32,35)^2 + (19-32,35)^2 + (42-32,35)^2$
$+ (56-32,35)^2 + (56-32,35)^2 + (44-32,35)^2 + (64-32,35)^2 + (17-32,35)^2$
$+ (17-32,35)^2] \cdot \frac{1}{21}$
$\rightarrow Var(x_3) = 314,807 \Rightarrow \sigma = \sqrt{314,807} \rightarrow \therefore \sigma = 17,742$ //
Moda: El valor que mas se repite
Para X_1
$Mo_1 = 1$ //
6 repeticiones
Para X_2
$Mo_2 = 3$ //
5 repeticiones
Para X_3
$Mo_3 = 17$ //
3 repeticiones

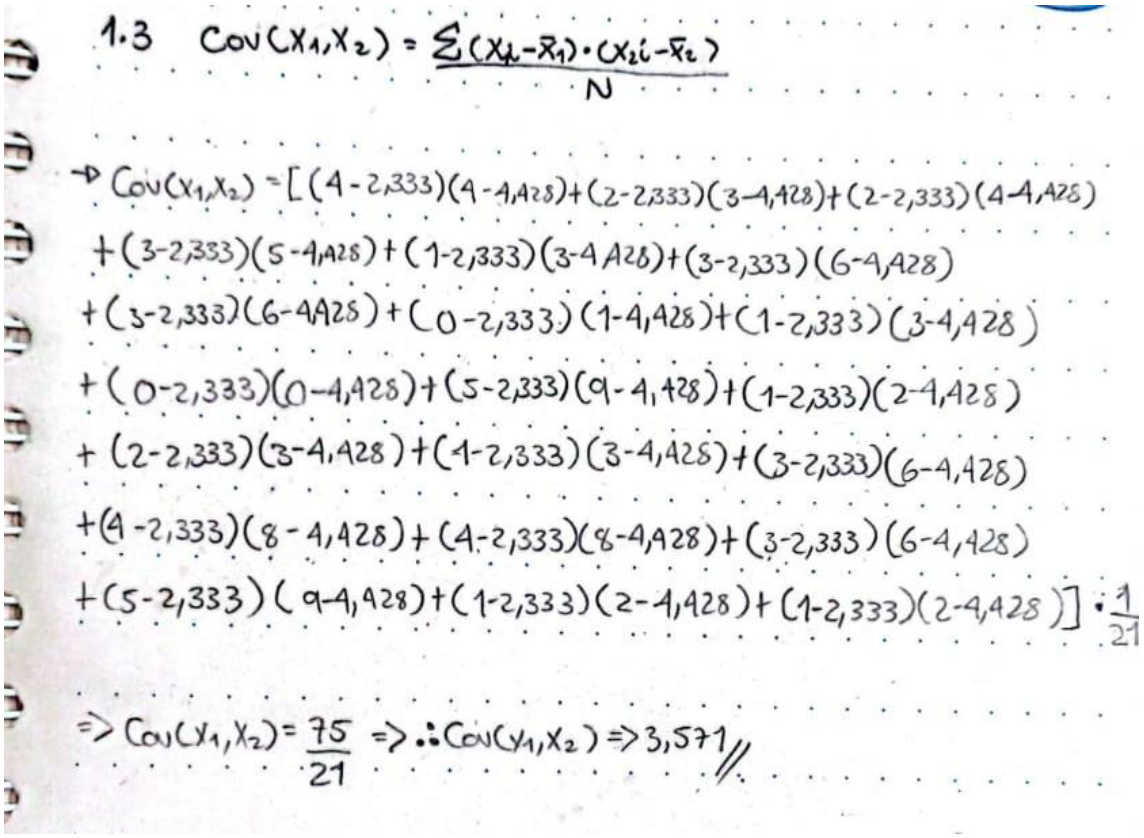
1.2.Dibujar un boxplot a mano





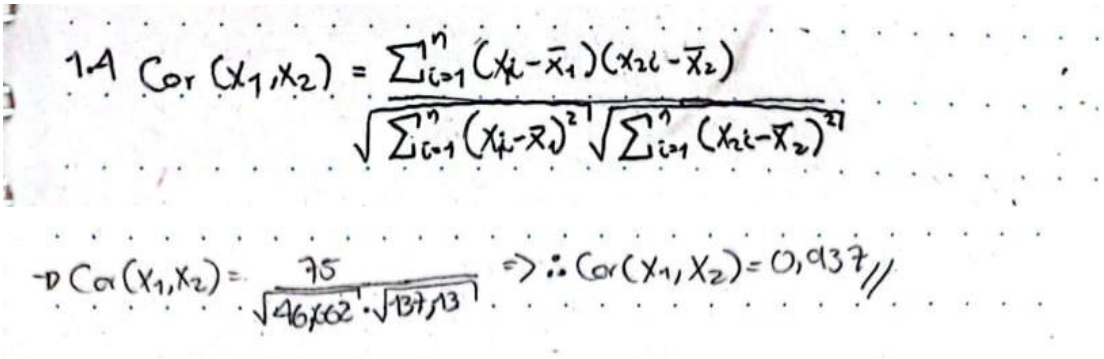
1.3.Cuál es la covarianza entre las 2 variables X1, X2

$$Cov(x,y) = \frac{\Sigma (x_i - \overline{x}) * (y_i - \overline{y})}{N}$$



1.4.Cuál es la correlación entre la variable x1 y x2 (Calcularla a mano).Correlación puede ser escrita también como:

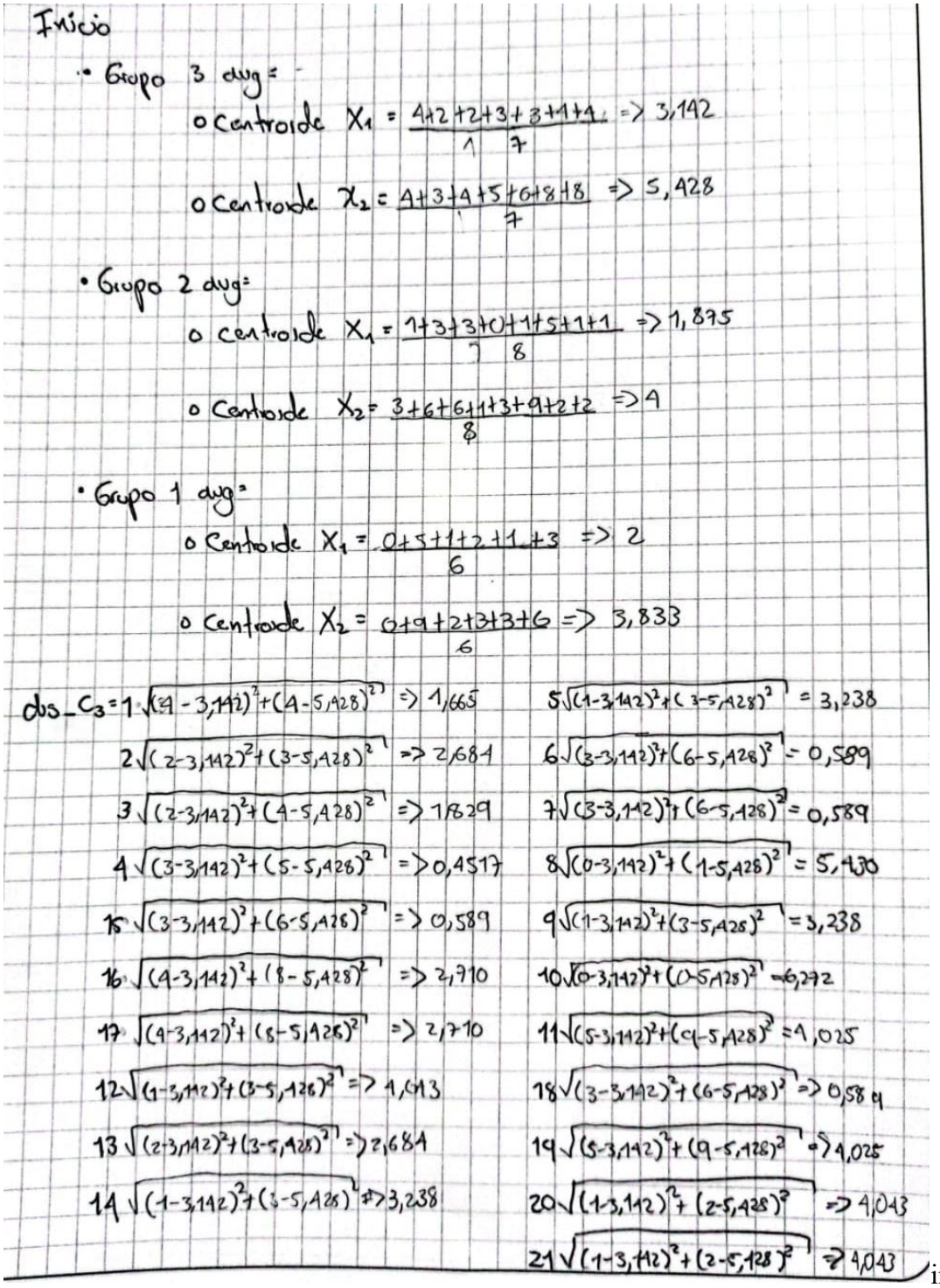
$$\mathrm{Cor}(X,Y) = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}},$$



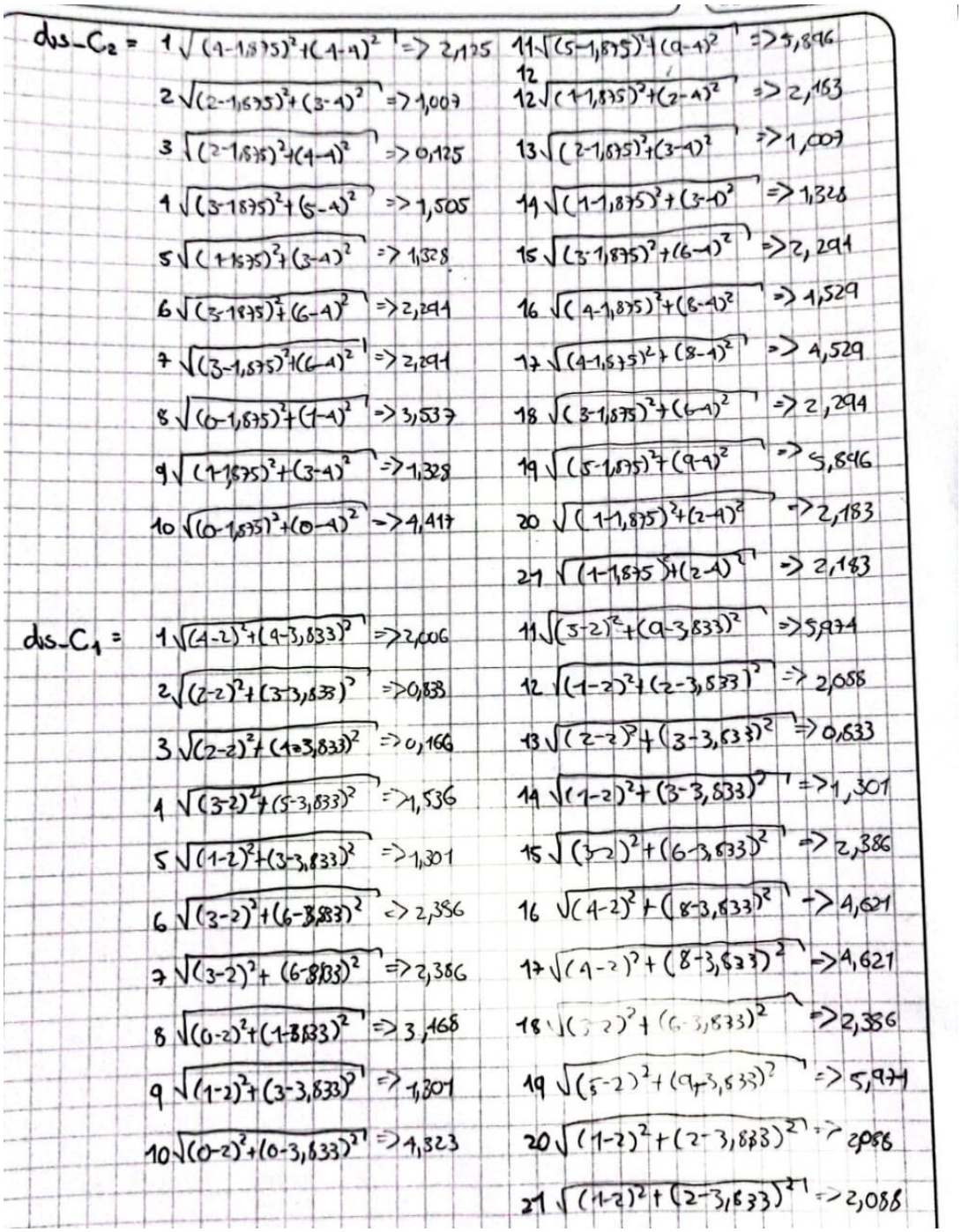
1.5. Explica la relación entre covarianza y correlación.

La covarianza explica que tipo de comportamiento tienen dos variables en conjunto, es decir, si el resultado es positivo quiere decir que si una aumenta la otra tiende a aumentar; si el resultado es negativo quiere decir que si una aumenta la otra tiende a disminuir.
Como se puede observar, su descripción concuerda por mucho con la correlación ya que también analiza como es el comportamiento de una variable respecto a la otra, en otras palabras, si existe una dependencia de una variable hacia la otra.
En lo único que difiere es en la interpretación, ya que la covarianza no esta normalizada, lo que puede dificultar la interpretación de esta y comparación entre las diferentes escalas de las variables, cosa que no pasa en la correlación, ya que esta última si esta normalizada (porque se está dividiendo básicamente por sus respectivas varianzas), es decir que sus datos varían entre -1 y 1, lo que ayuda significativamente a interpretar los valores.

1.6. Calcule el resultado del algoritmo K-means sobre este set de datos a mano como lo hicimos en Excel. Vamos a crear 3 grupos, es decir K=3



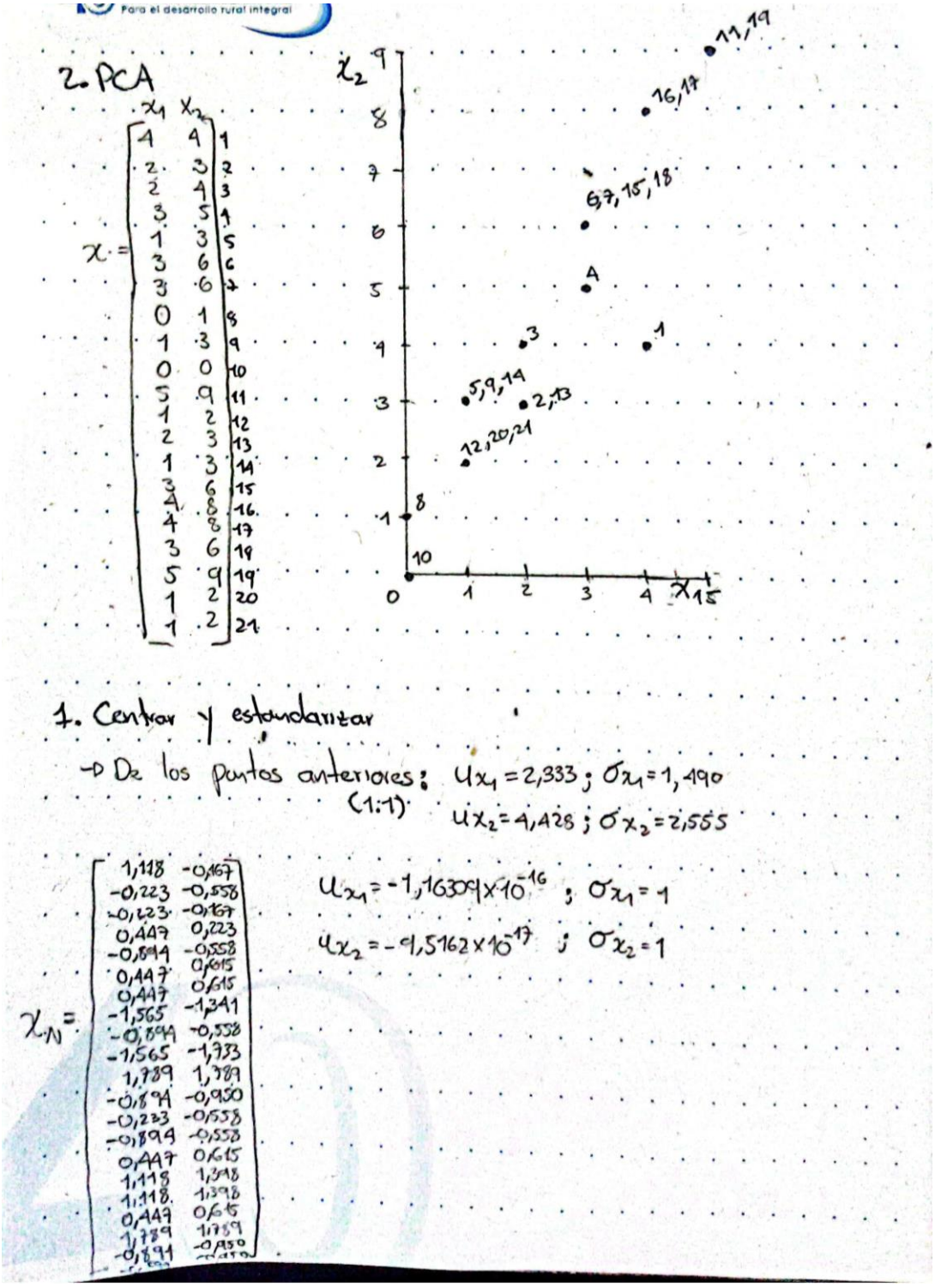
ir, k=3 (clústeres).



Siguiendo la lógica con cada iteración hasta que no haya cambios en las etiquetas tenemos:

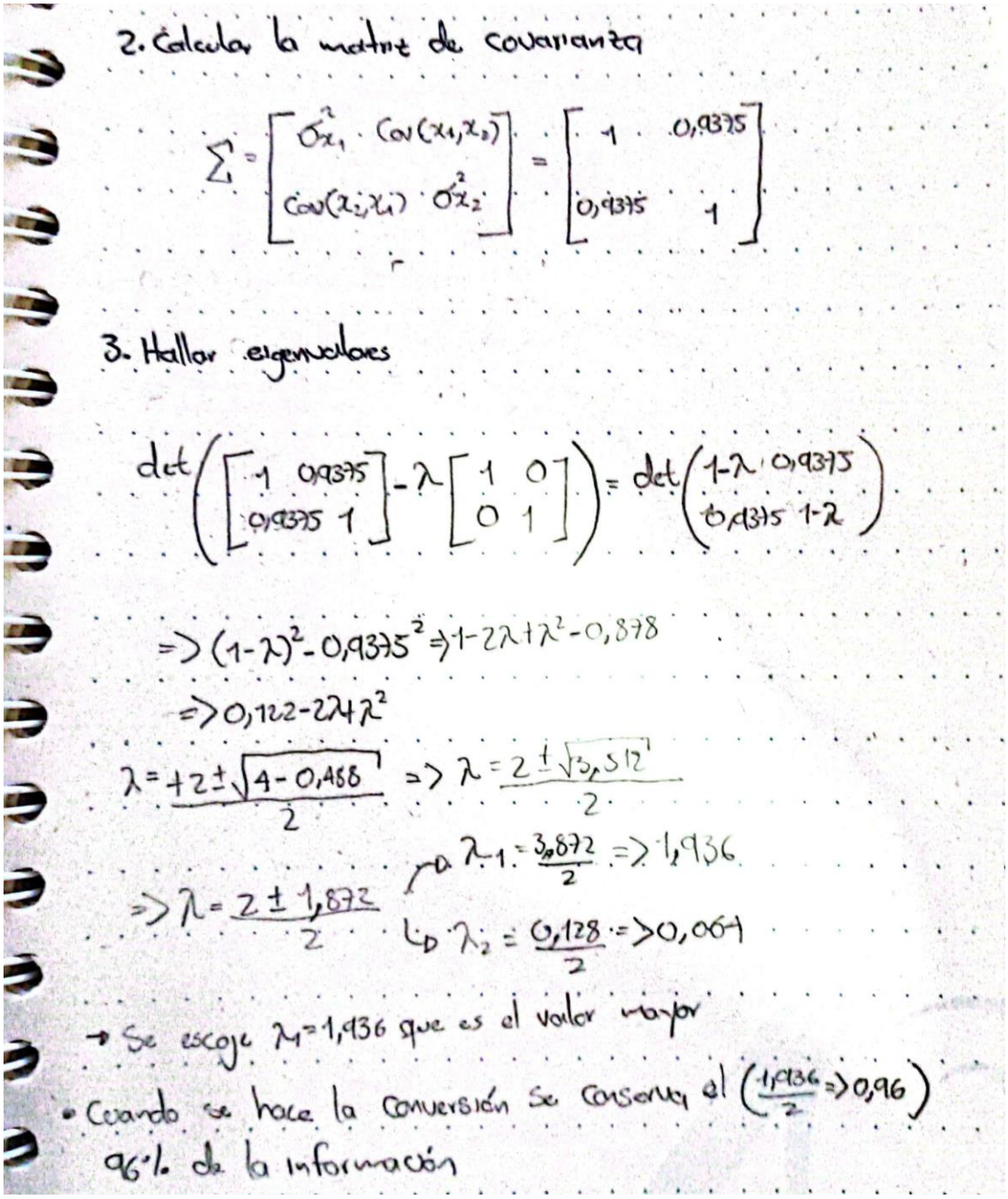
	x1	x2	Inicio	centroide x1	centroide x2	dis_c3_3	dis_c2_2	dis_c1_1	New_labels	centroide x1_it2	centroide x2_it2	dis_c3_3_it2	dis_c2_2_it2	dis_c1_1_it2	New_labels	centroide x1_it3	centroide x2_it3	dis_c3_3_it3	dis_c2_2_it3	dis_c1_1_it3	New_labels
1	4	4	3	*3,14285714*	*5,428571429*	1,66598626	2,125	2,00693243	3	*3,7*	*6,7*	2,71661554	2	3,49857114	2	*3,333333333*	*6,444444444*	2,533723167	1,414213562	3,816084381	2
2	2	3	3			2,68404203	1,00778222	0,83333333	1			4,07185461	1	1,28062485	2			3,693504475	1	1,600781059	2
3	2	4	3			1,82946407	0,125	0,16666667	2			3,19061123	0	2,05912603	2			2,784436464	0,632455532	2,358495283	2
4	3	5	3			0,45175395	1,50519932	1,53659074	3			1,83847763	1,41421356	3,44093011	2			1,482407118	1,264911064	3,75	2
5	1	3	2	*1,875*	*4*	3,23879544	1,32876823	1,30170828	1	*2*	*4*	4,580393	1,41421356	0,8	1	*2,6*	*3,8*	4,160365606	1,788854382	1,030776406	1
6	3	6	2			0,58901509	2,29469497	2,38630351	3			0,98994949	2,23606798	4,29418211	3			0,555555556	2,236067977	4,589389938	3
7	3	6	2			0,58901509	2,29469497	2,38630351	3			0,98994949	2,23606798	4,29418211	3			0,555555556	2,236067977	4,589389938	3
8	0	1	2			5,4304508	3,53774292	3,46810867	1			6,7955868	3,60555128	1,56204994	1			6,38381441	3,820994635	1,25	1
9	1	3	2			3,23879544	1,32876823	1,30170828	1			4,580393	1,41421356	0,8	1			4,160365606	1,788854382	1,030776406	1
10	0	0	1	*2*	*3,833333333*	6,27271383	4,41764926	4,32370726	1	*1*	*2,2*	7,65375725	4,47213595	2,41660919	1	*0,75*	*2*	7,255478985	4,604345773	2,136000936	1
11	5	9	1			4,02542937	5,89623821	5,97448278	3			2,64196896	5,83095189	7,88923317	3			3,051006715	5,727128425	8,189169677	3
12	1	2	1			4,04313477	2,18303115	2,08832735	1			5,42033209	2,23606798	0,2	1			5,019714221	2,408318916	0,25	1
13	2	3	1			2,68404203	1,00778222	0,83333333	1			4,07185461	1	1,28062485	2			3,693504475	1	1,600781059	2
14	1	3	1			3,23879544	1,32876823	1,30170828	1			4,580393	1,41421356	0,8	1			4,160365606	1,788854382	1,030776406	1
15	3	6	3			0,58901509	2,29469497	2,38630351	3			0,98994949	2,23606798	4,29418211	3			0,555555556	2,236067977	4,589389938	3
16	4	8	3			2,71052371	4,52941773	4,62180821	3			1,33416641	4,47213595	6,52993109	3			1,692394024	4,427188724	6,823672032	3
17	4	8	3			2,71052371	4,52941773	4,62180821	3			1,33416641	4,47213595	6,52993109	3			1,692394024	4,427188724	6,823672032	3
18	3	6	1			0,58901509	2,29469497	2,38630351	3			0,98994949	2,23606798	4,29418211	3			0,555555556	2,236067977	4,589389938	3
19	5	9	2			4,02542937	5,89623821	5,97448278	3			2,64196896	5,83095189	7,88923317	3			3,051006715	5,727128425	8,189169677	3
20	1	2	2			4,04313477	2,18303115	2,08832735	1			5,42033209	2,23606798	0,2	1			5,019714221	2,408318916	0,25	1
21	1	2	2			4,04313477	2,18303115	2,08832735	1			5,42033209	2,23606798	0,2	1			5,019714221	2,408318916	0,25	1

2. PCA. Utilizar los datos de la tabla 1, para calcular PCA y reducir la dimensionalidad de 2 dimensiones a 1. Para este ejercicio se debe utilizar las variables X1, y X2 y crear un vector con una sola dimensión.

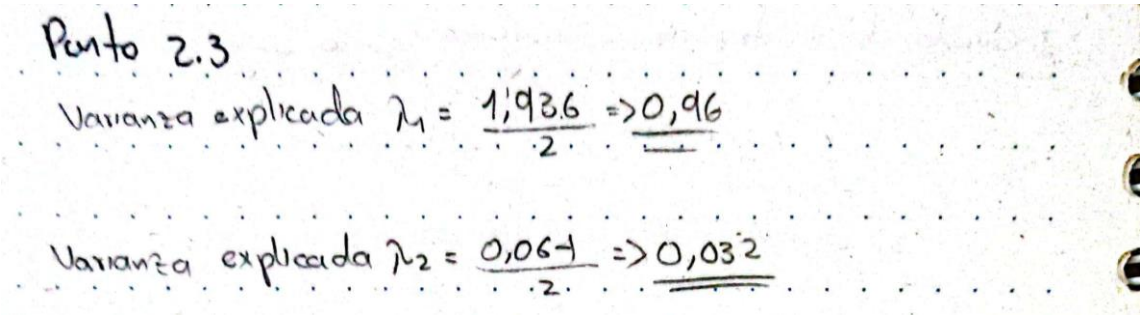


2.1. Cuál es la matriz de covarianza

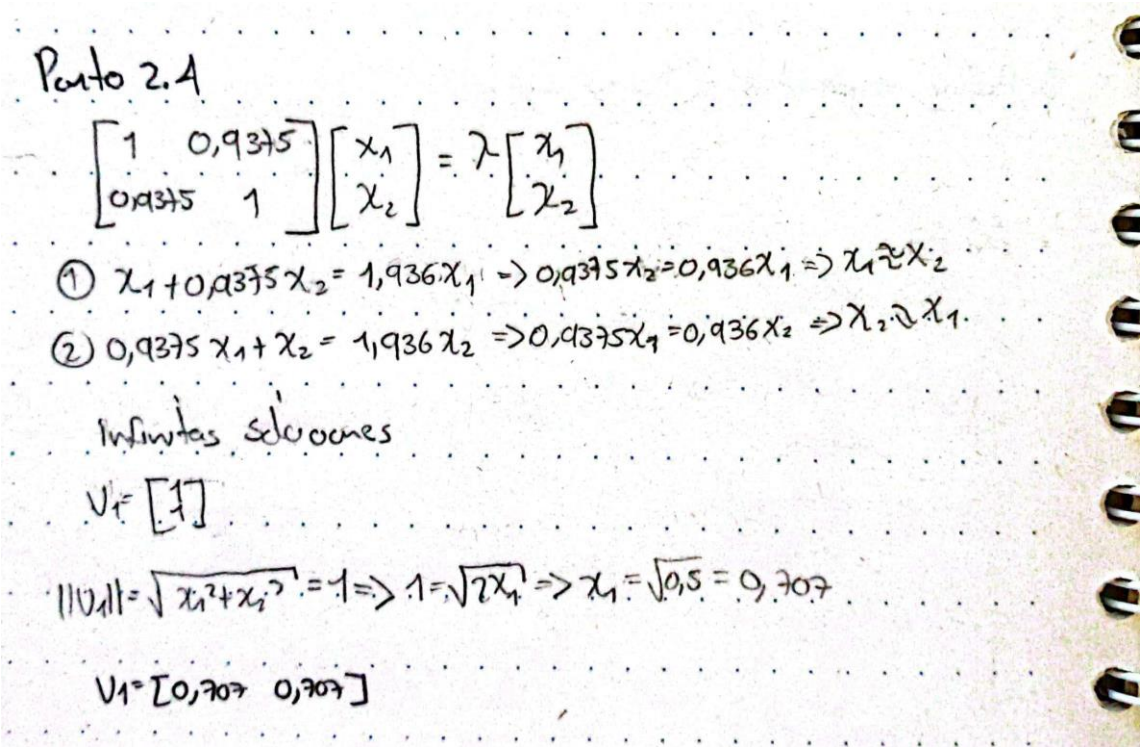
2.2. Cuáles son los eigenvalores



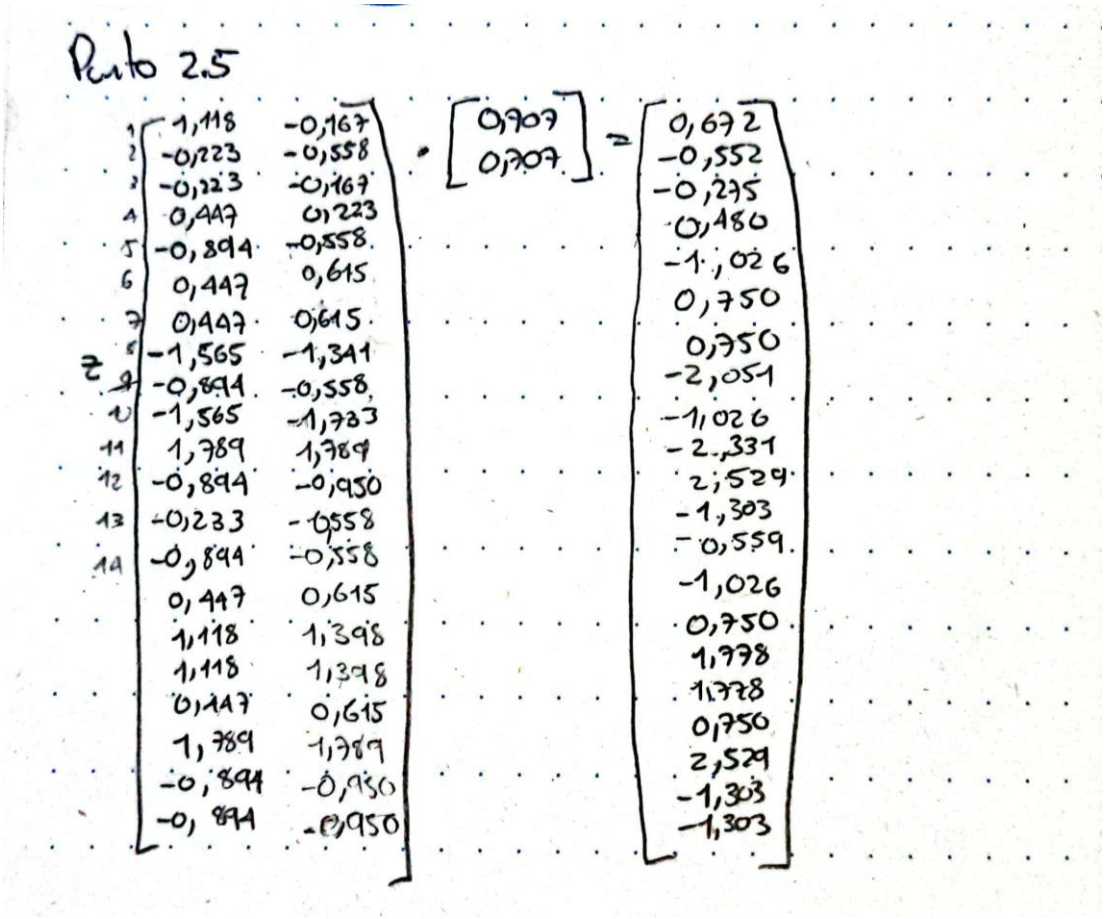
2.3. Cuál es la varianza explicada por el eigenvalor.



2.4. Cuál es el valor del eigenvector



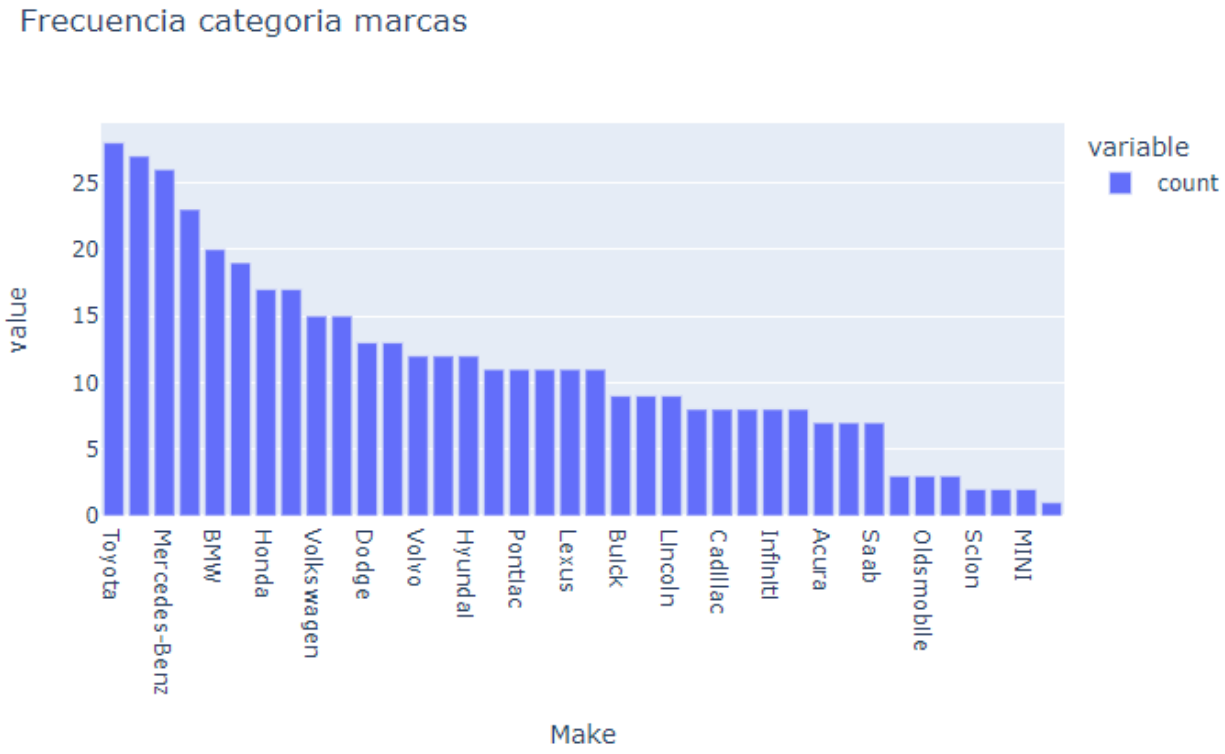
2.5. Cuál es la matriz proyectada.

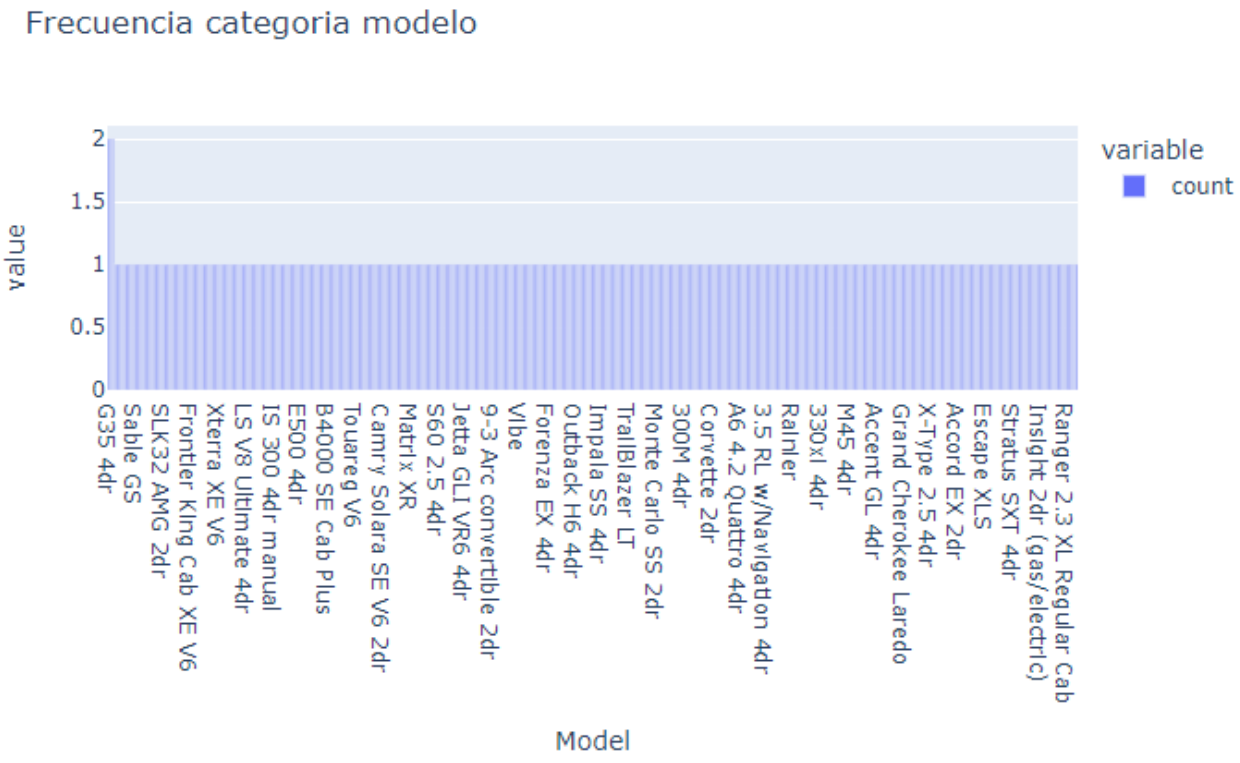


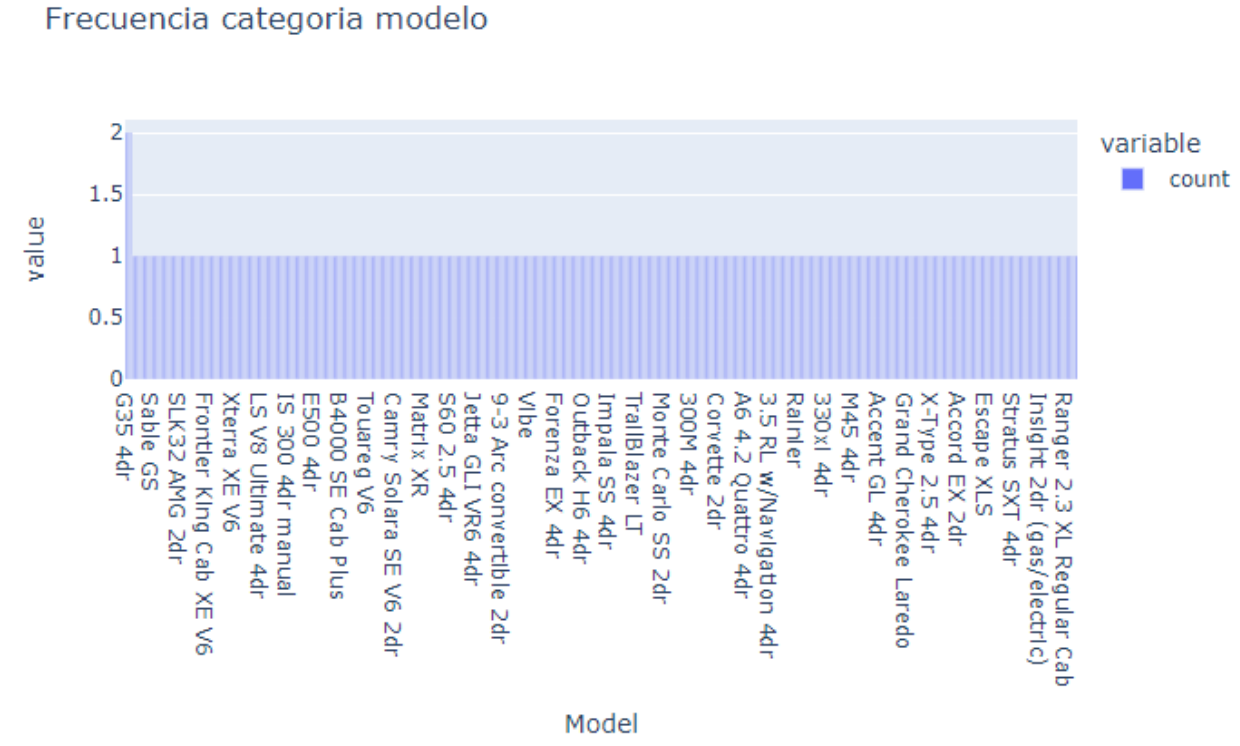
3. Utilizando el dataset del proyecto data/CARS.csv crear: Utilizar la librería de plotly.

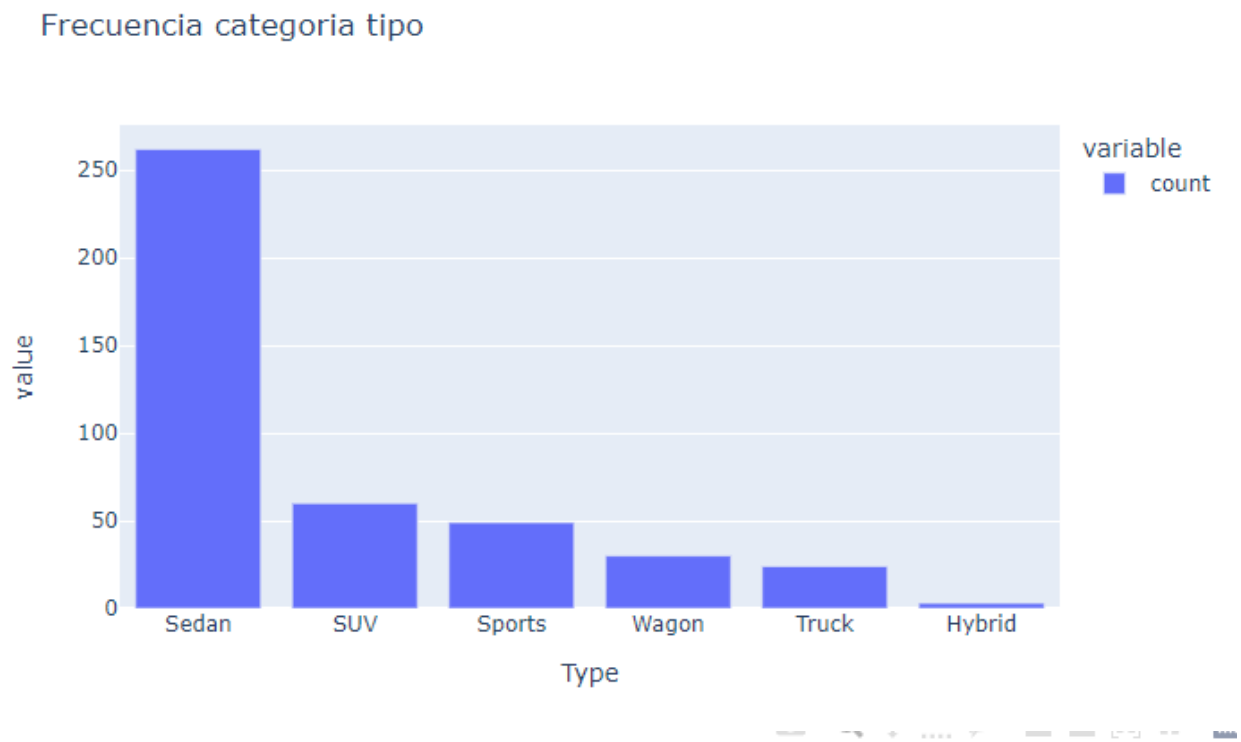
3.1. Distribución de cada variable:

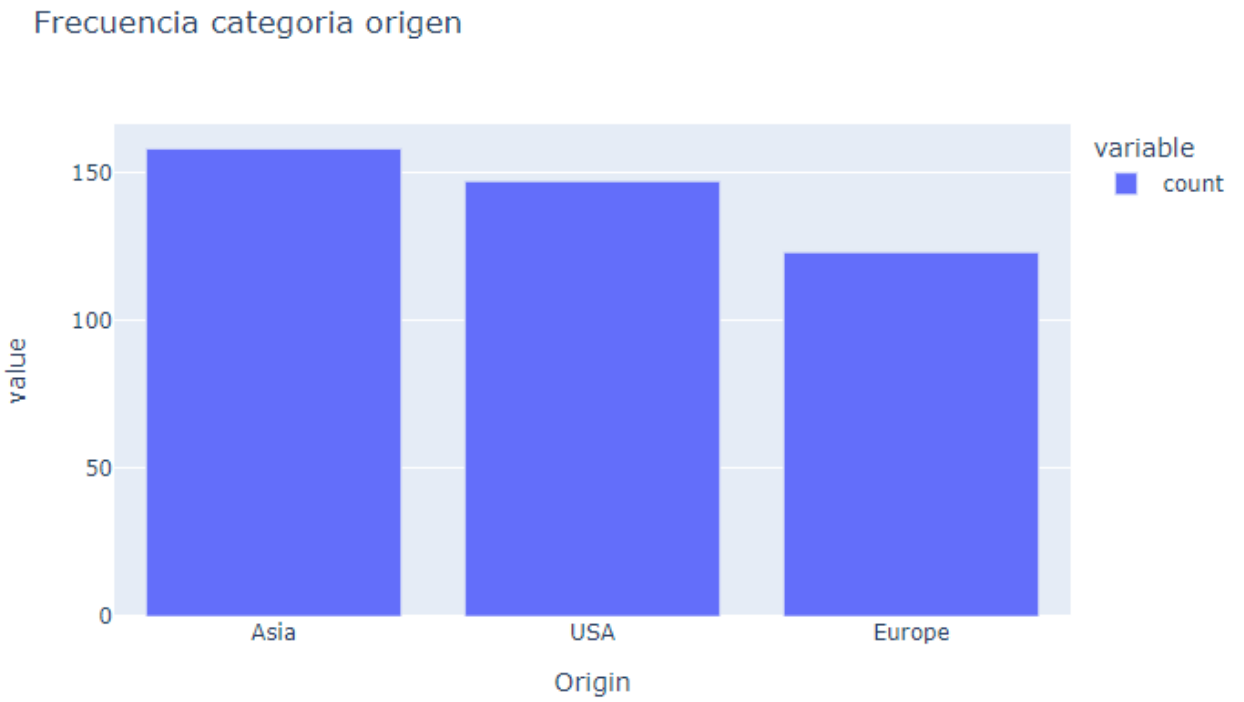
3.1.1. Para las variables categóricas un gráfico de barras. Categoría número de observaciones.

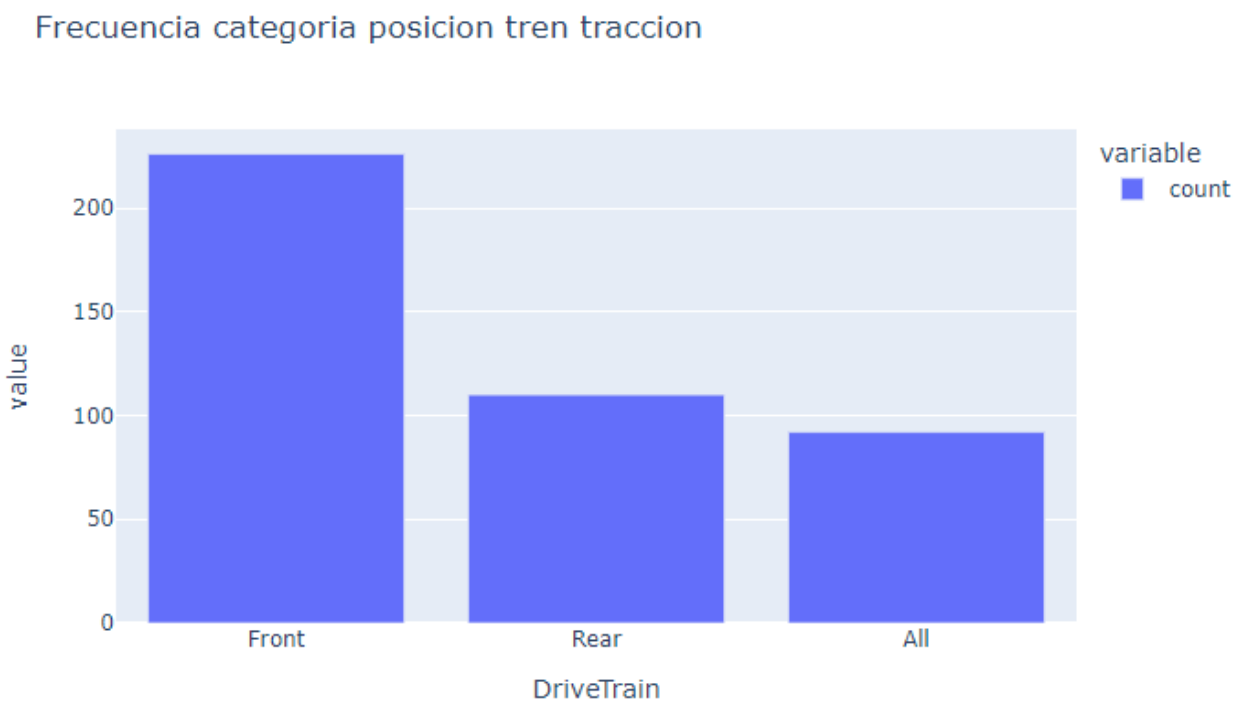



Frecuencia categoria modelo
variable
count
value
2
1.5
1
0.5
0
Ranger 2.3 XL Regular Cab
Insight 2dr (gas/electric)
Stratus SXT 4dr
Escape XLS
Accord EX 2dr
X-Type 2.5 4dr
Grand Cherokee Laredo
Accent GL 4dr
M45 4dr
330xi 4dr
Rainier
3.5 RL w/Navigation 4dr
A6 4.2 Quattro 4dr
Corvette 2dr
300M 4dr
Monte Carlo SS 2dr
TrailBlazer LT
Impala SS 4dr
Outback H6 4dr
Forenza EX 4dr
Vibe
9-3 Arc convertible 2dr
Jetta GLI VR6 4dr
S60 2.5 4dr
Matrix XR
Camry Solara SE V6 2dr
Touareg V6
B4000 SE Cab Plus
E500 4dr
IS 300 4dr manual
LS V8 Ultimate 4dr
Xterra XE V6
Frontier King Cab XE V6
SLK32 AMG 2dr
Sable GS
G35 4dr
Model


Frecuencia categoria modelo
variable
count
value
2
1.5
1
0.5
0
Ranger 2.3 XL Regular Cab
Insight 2dr (gas/electric)
Stratus SXT 4dr
Escape XLS
Accord EX 2dr
X-Type 2.5 4dr
Grand Cherokee Laredo
Accent GL 4dr
M45 4dr
330xi 4dr
Rainier
3.5 RL w/Navigation 4dr
A6 4.2 Quattro 4dr
Corvette 2dr
300M 4dr
Monte Carlo SS 2dr
TrailBlazer LT
Impala SS 4dr
Outback H6 4dr
Forenza EX 4dr
Vibe
9-3 Arc convertible 2dr
Jetta GLI VR6 4dr
S60 2.5 4dr
Matrix XR
Camry Solara SE V6 2dr
Touareg V6
B4000 SE Cab Plus
E500 4dr
IS 300 4dr manual
LS V8 Ultimate 4dr
Xterra XE V6
Frontier King Cab XE V6
SLK32 AMG 2dr
Sable GS
G35 4dr
Model

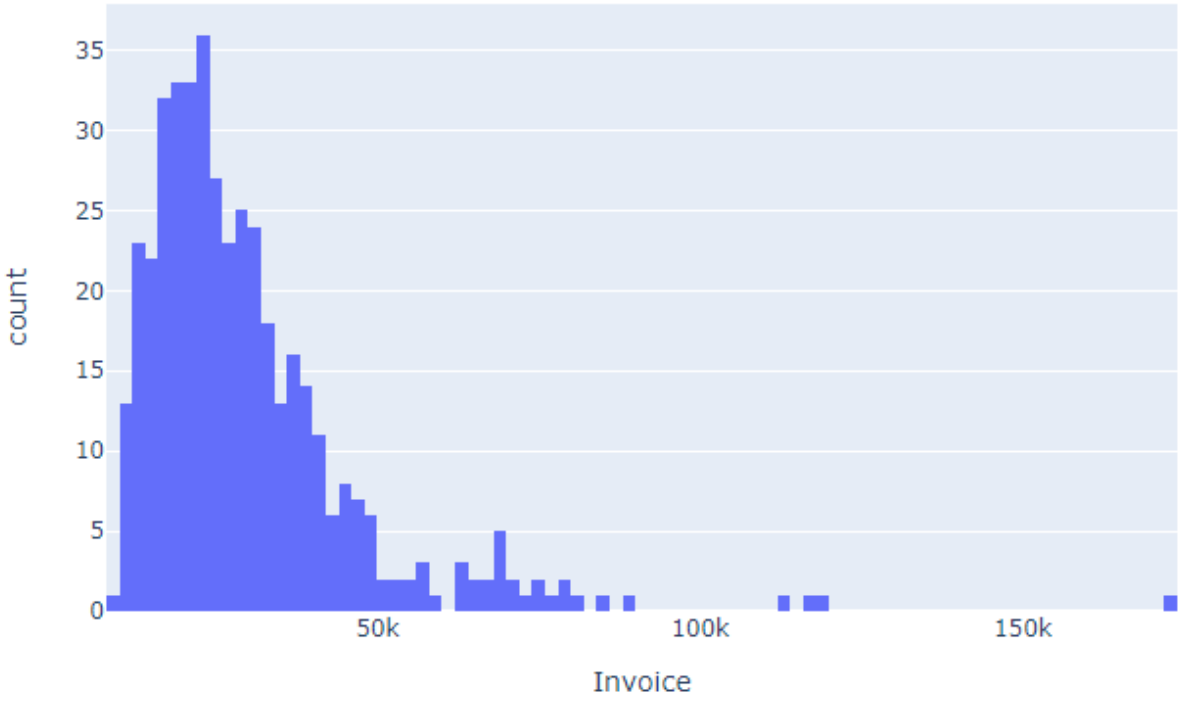

Frecuencia categoria tipo
variable
count
value
250
200
150
100
50
0
Sedan
SUV
Sports
Wagon
Truck
Hybrid
Type

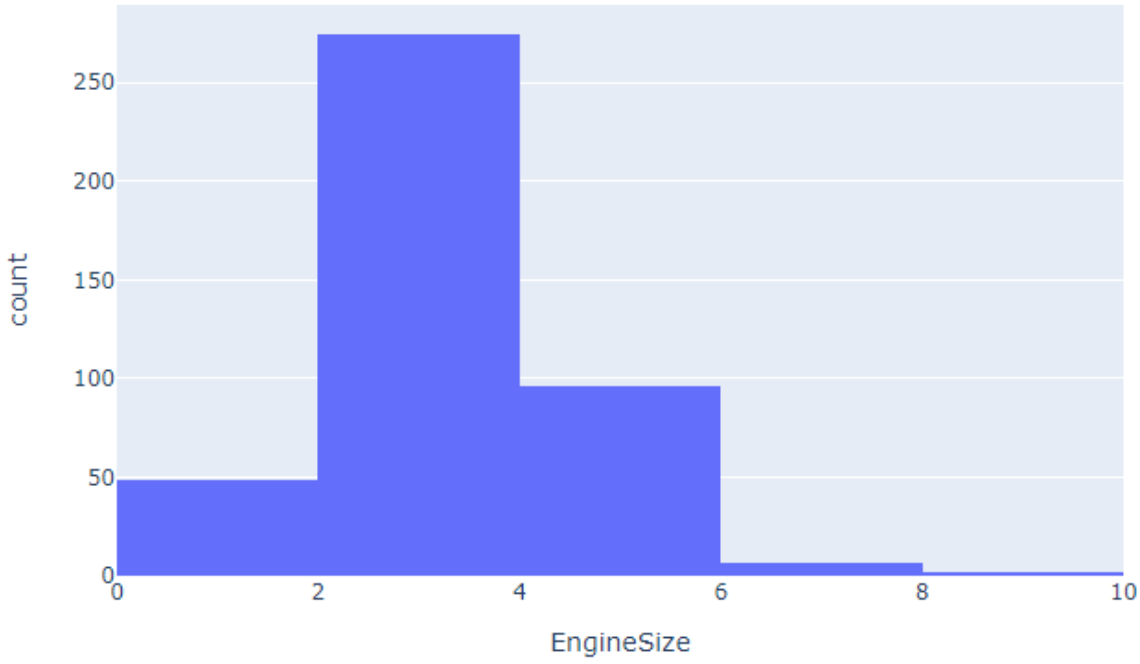

Frecuencia categoria origen
variable
count
value
150
100
50
0
Asia
USA
Europe
Origin

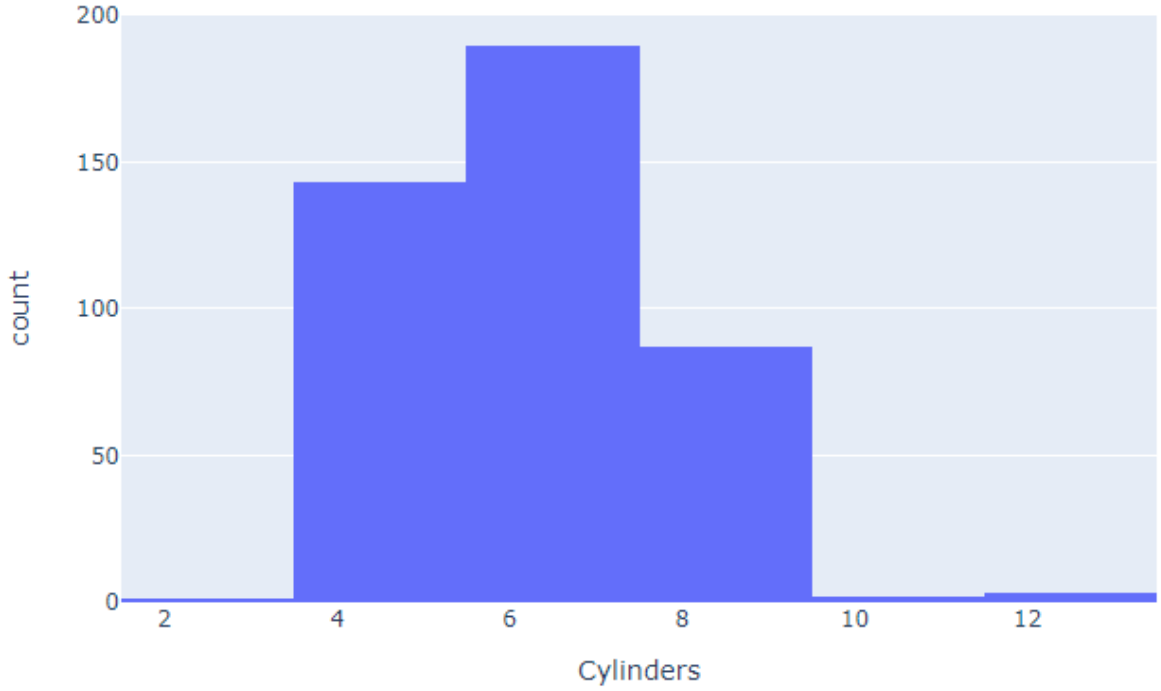


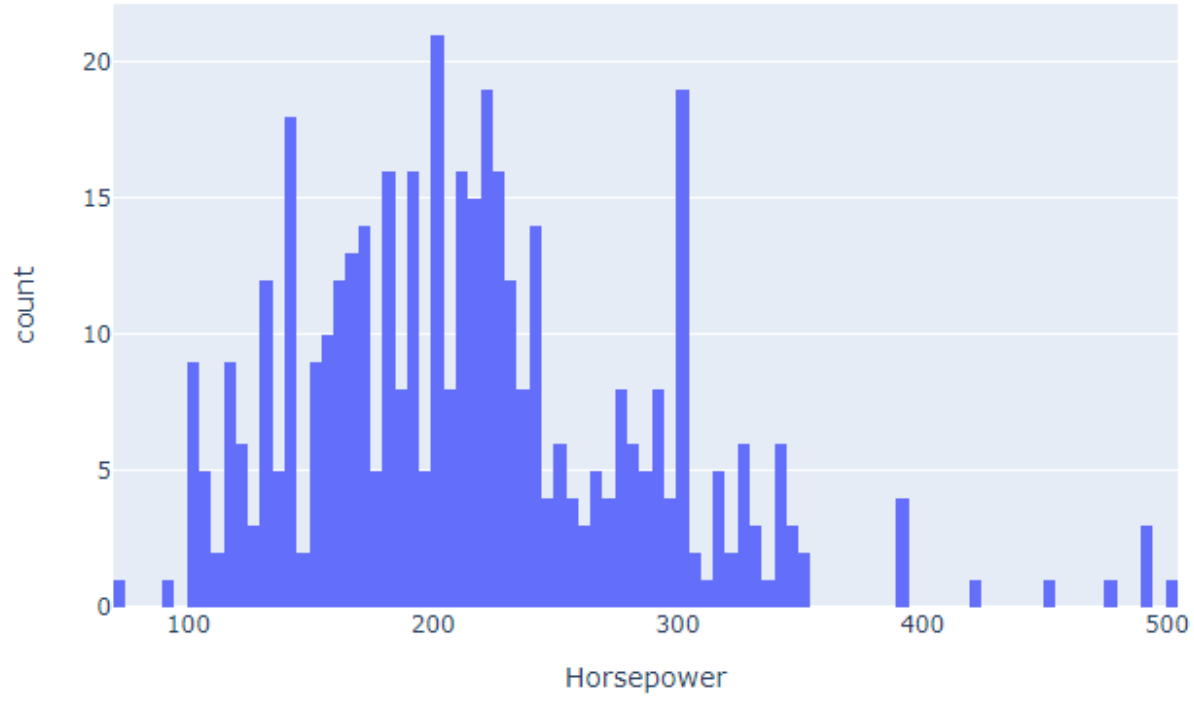
3.1.2. Para las variables numéricas crear histogramas. Listar los modelos de carros que están más lejos de 5 estándares de desviación, y serían considerados outliers. Hacer test de si es una distribución normal o no.

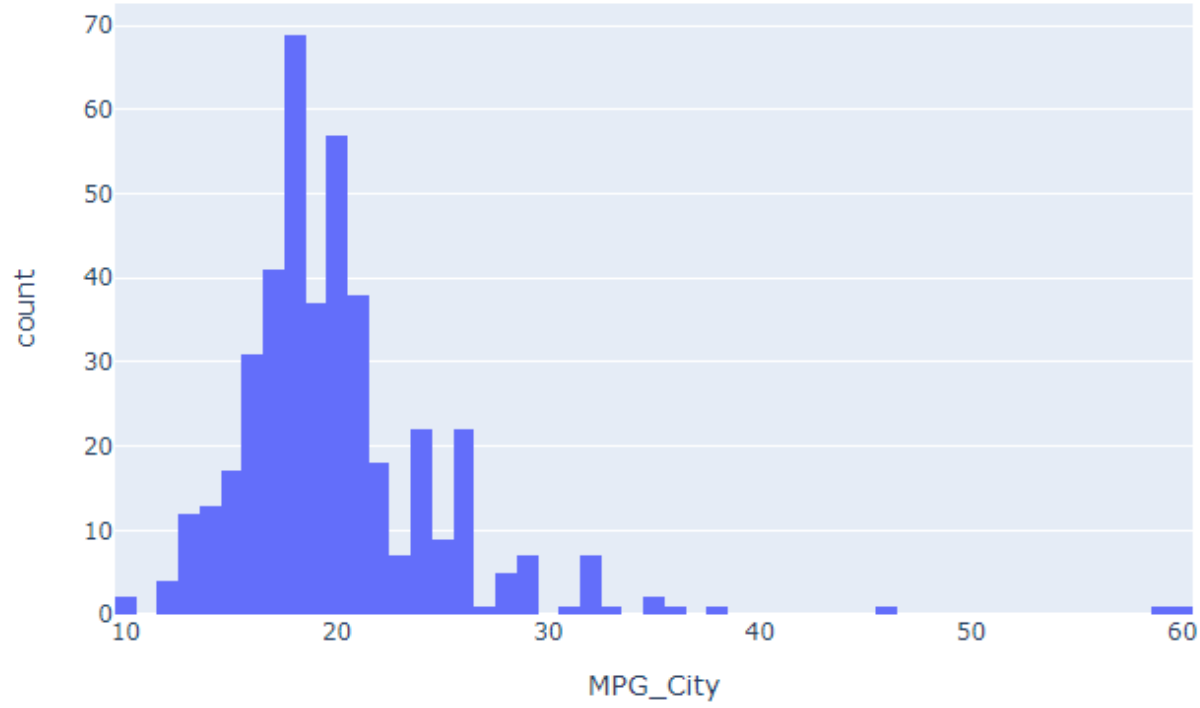
Histogramas:

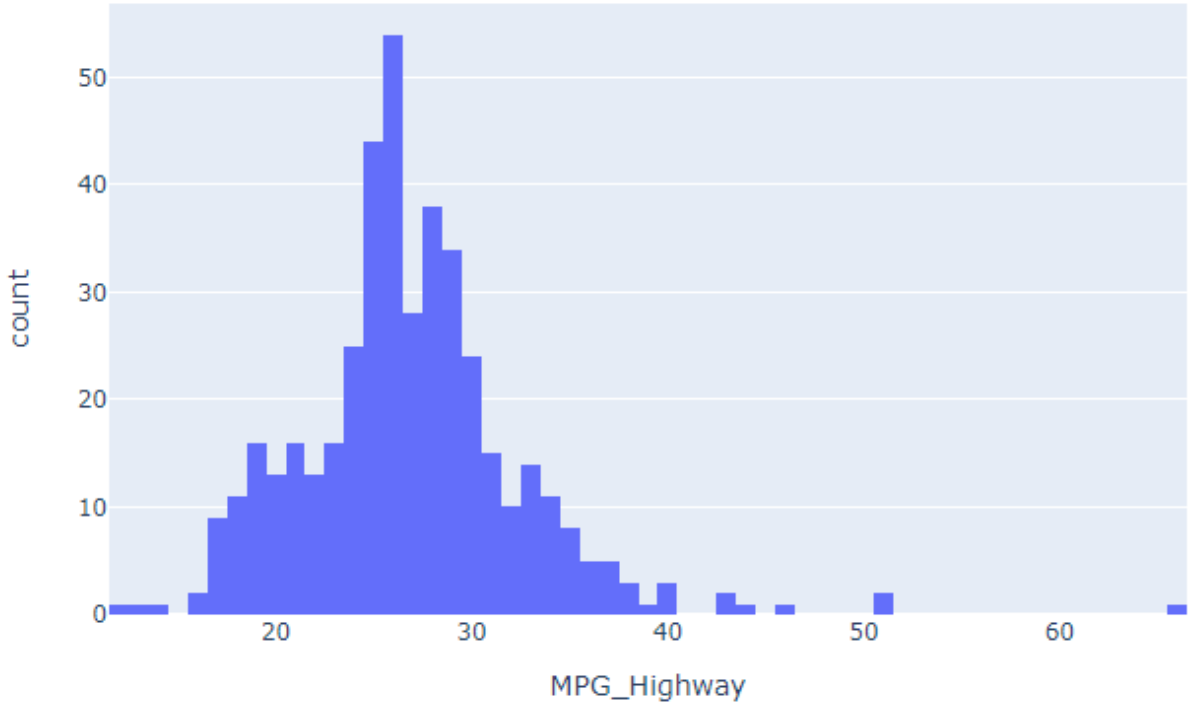

count
35
30
25
20
15
10
5
0
50k
100k
150k
Invoice

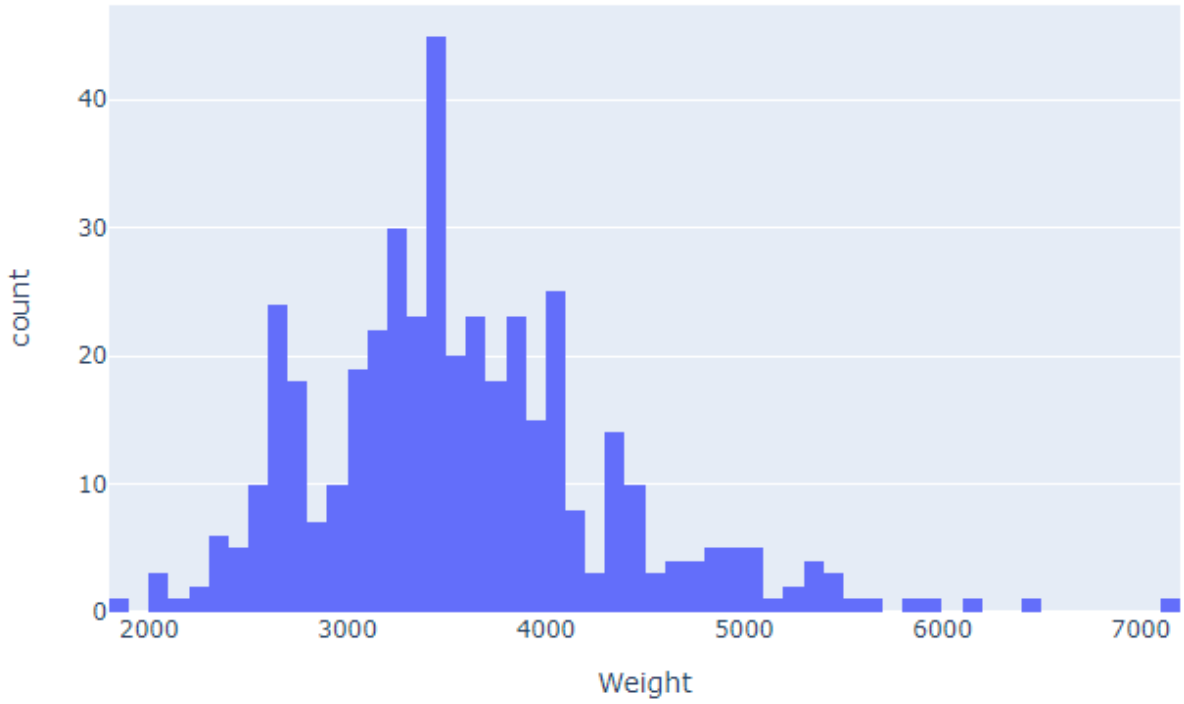

count
250
200
150
100
50
0
0
2
4
6
8
10
EngineSize

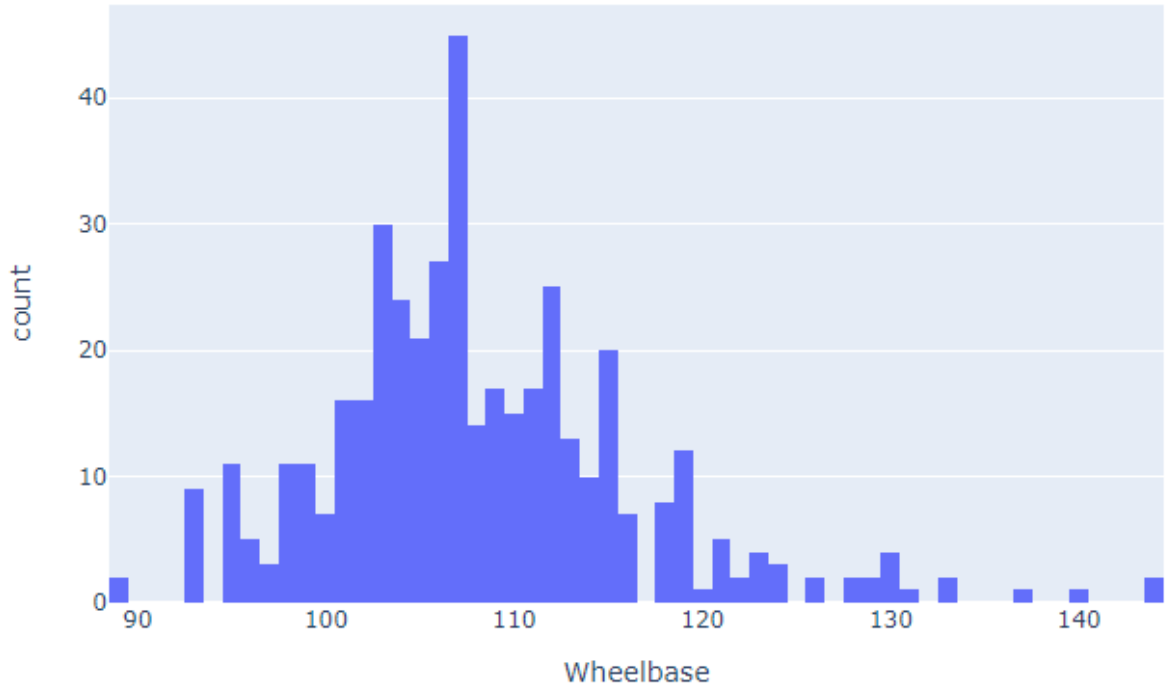

200
150
100
50
0
count
2
4
6
8
10
12
Cylinders

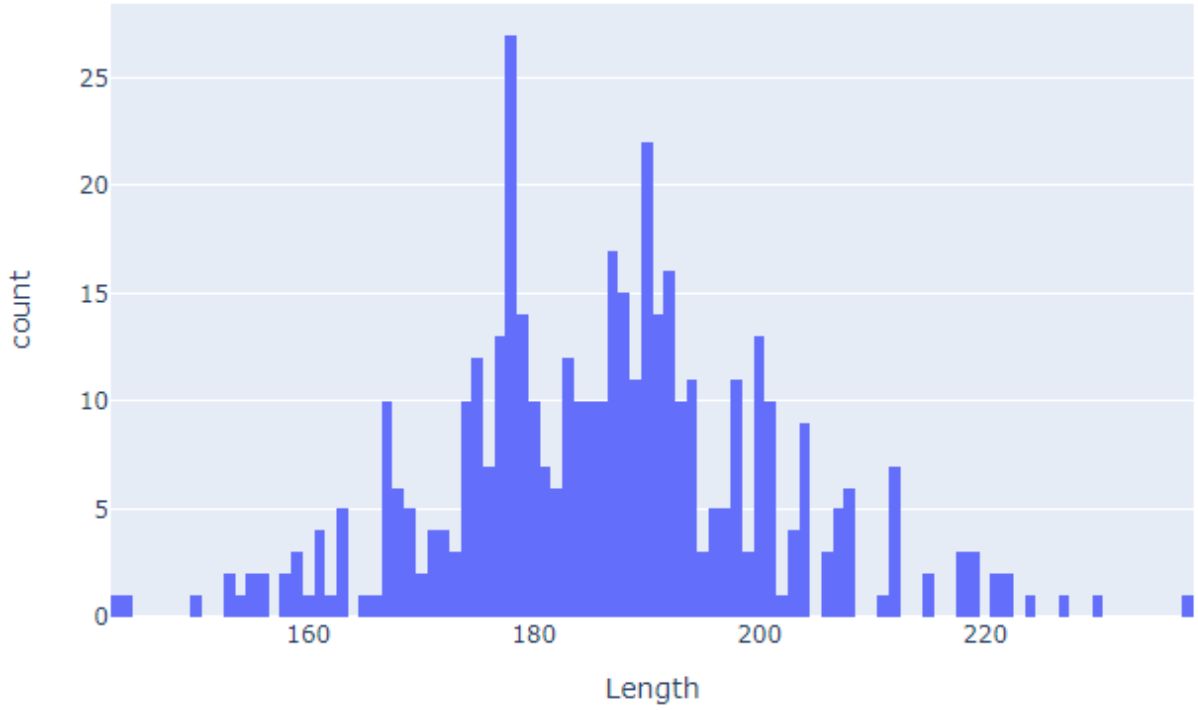

20
15
10
5
0
count
100
200
300
400
500
Horsepower


70
60
50
40
30
20
10
0
count
10
20
30
40
50
60
MPG_City

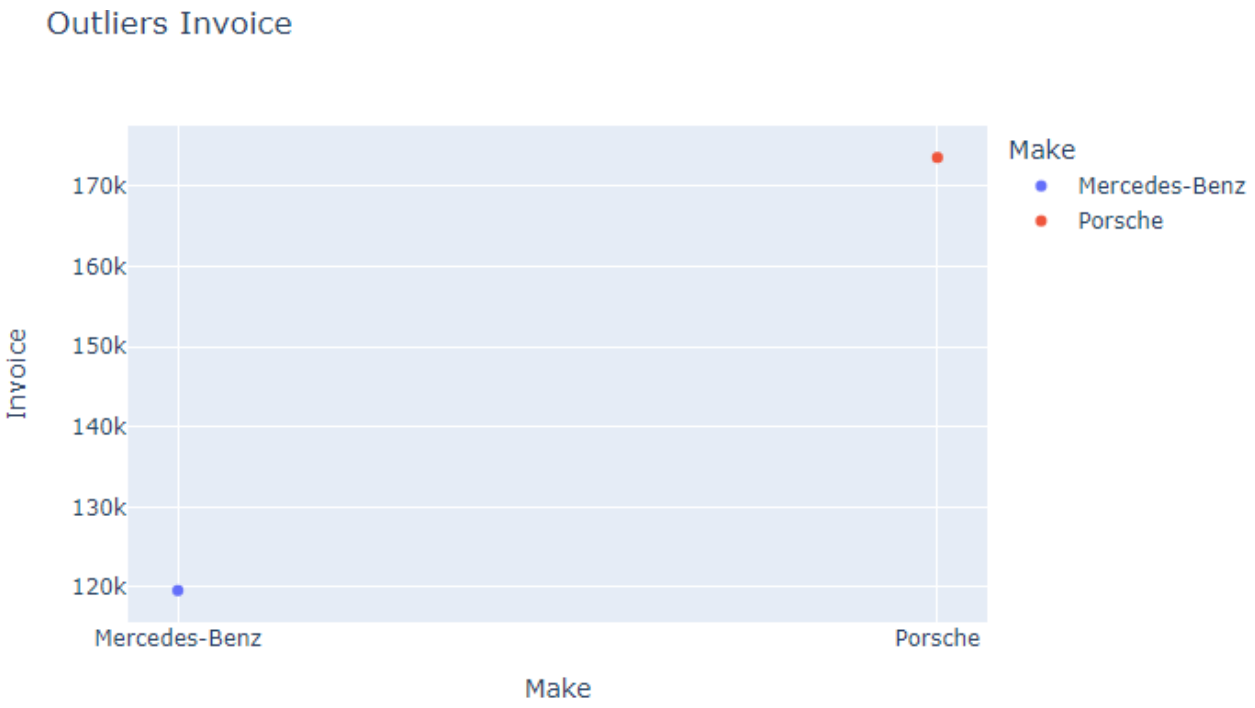

50
40
30
20
10
0
count
20
30
40
50
60
MPG_Highway

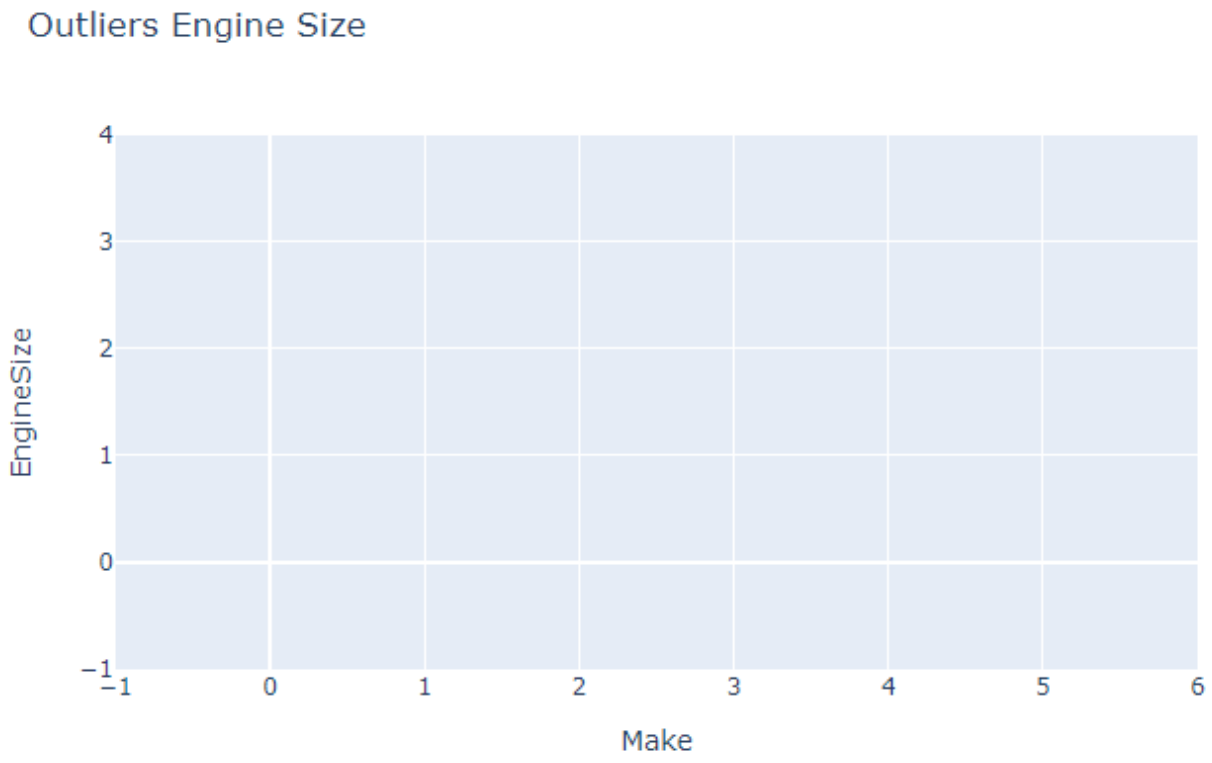

count
40
30
20
10
0
2000
3000
4000
5000
6000
7000
Weight

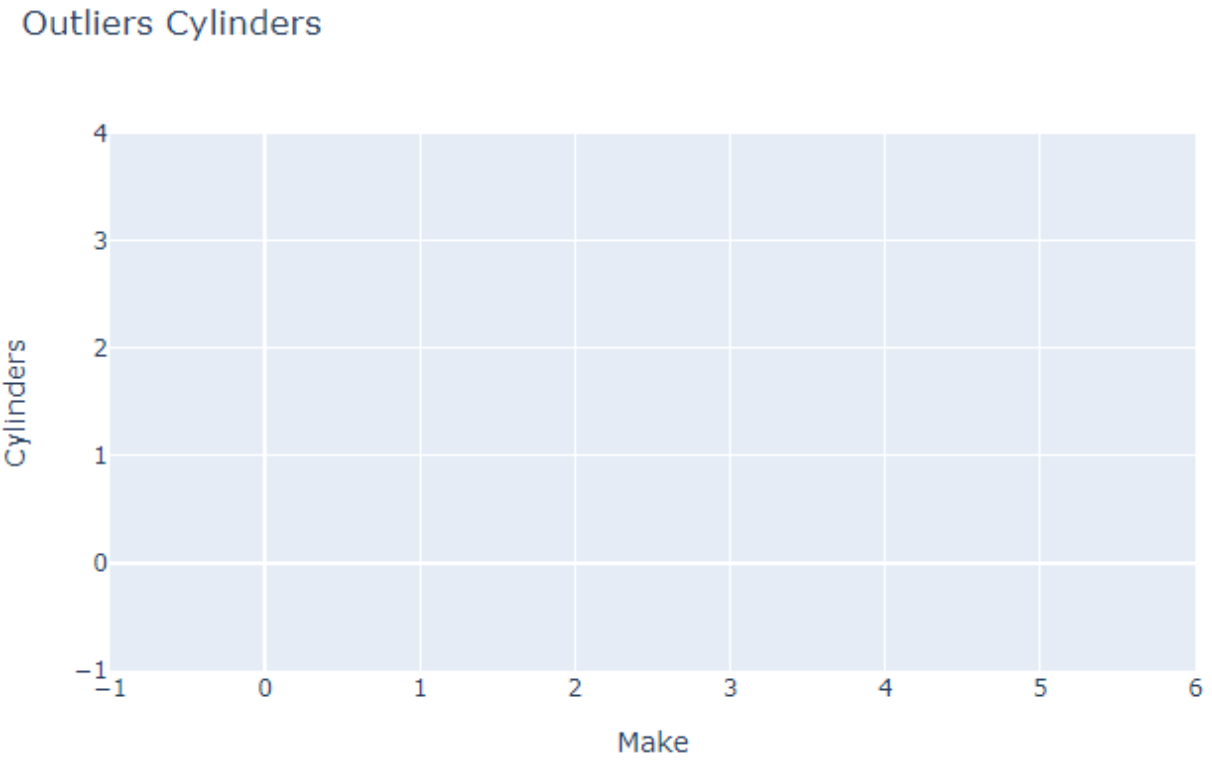

count
40
30
20
10
0
90
100
110
120
130
140
Wheelbase

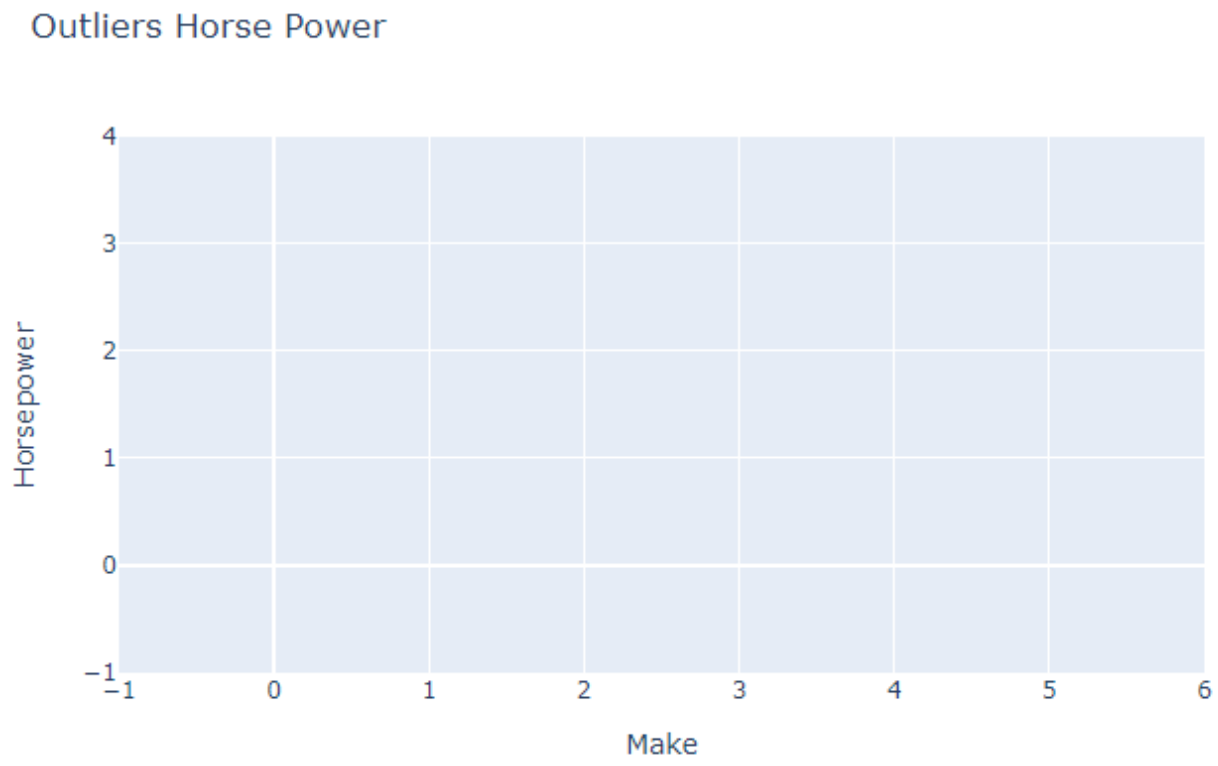


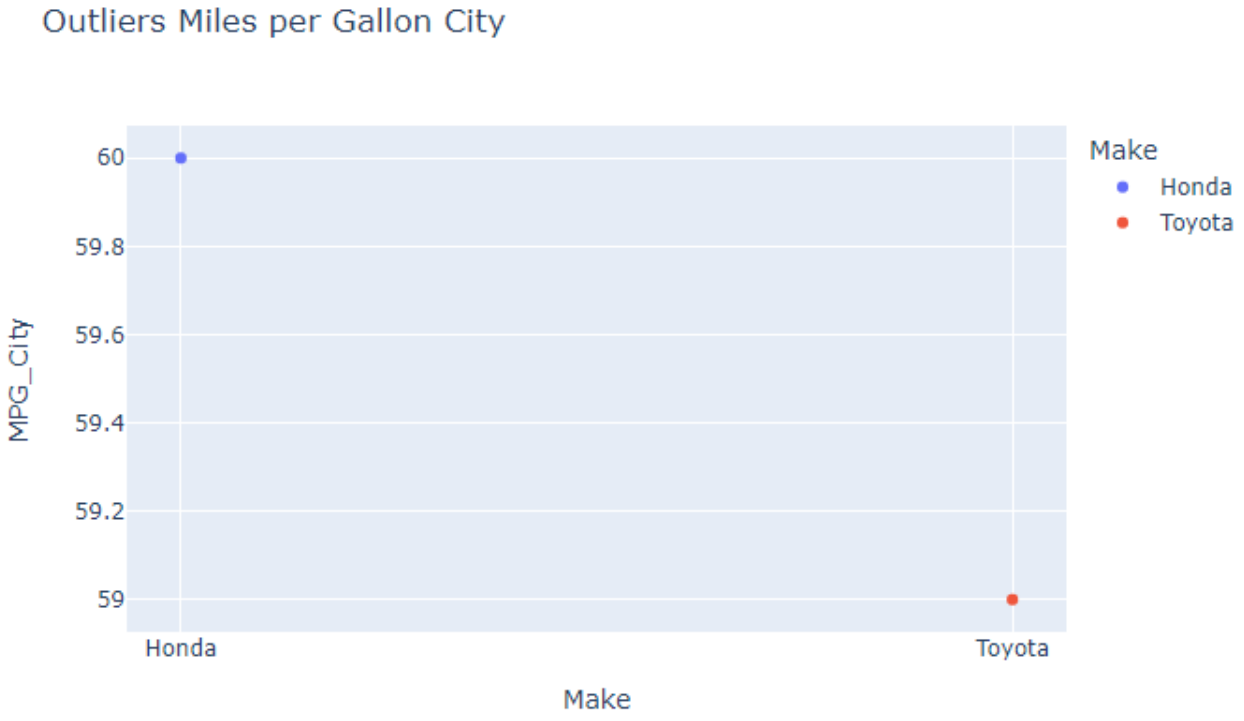
Análisis de Outliers 5 desviaciones estándar de diferencia:

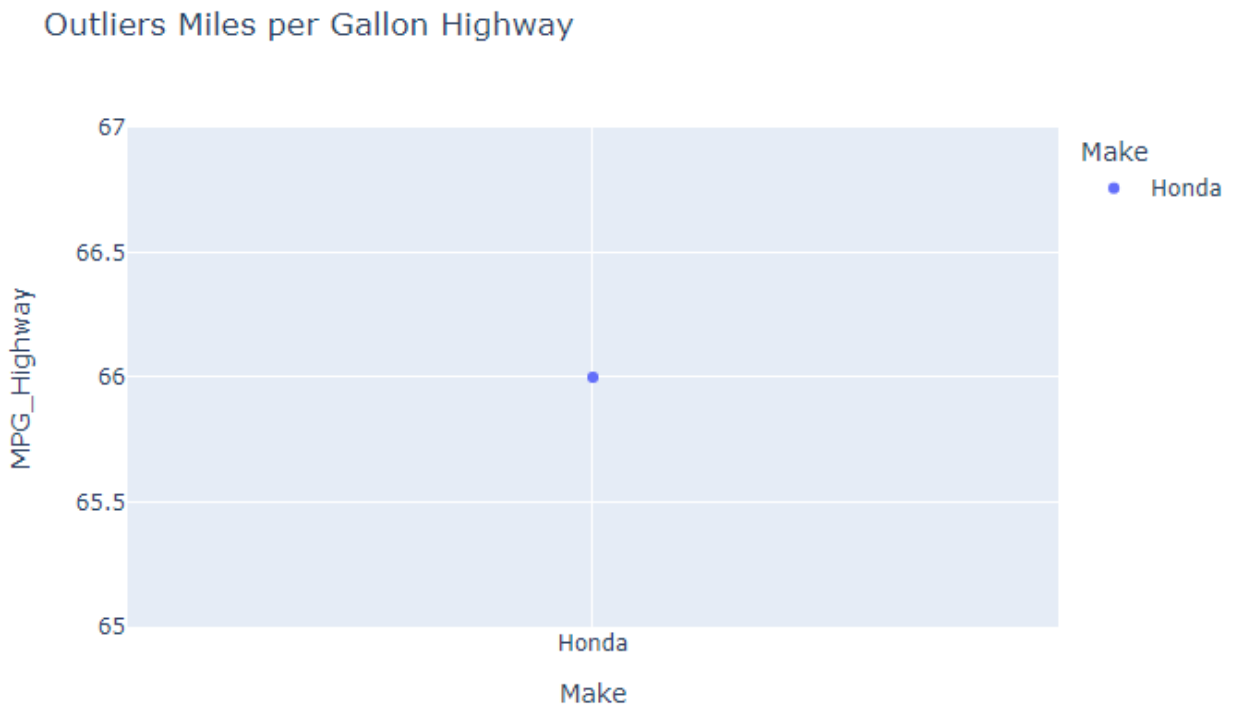


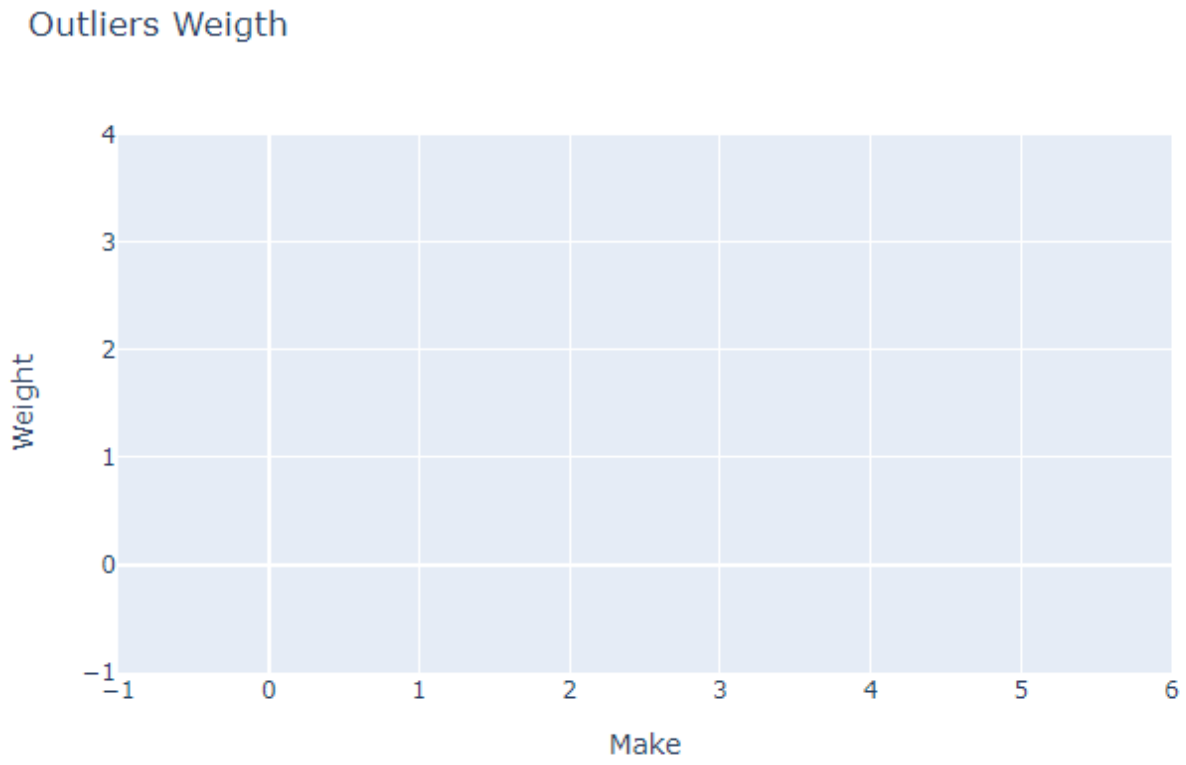

Outliers Engine Size
EngineSize
4
3
2
1
0
−1
−1
0
1
2
3
4
5
6
Make

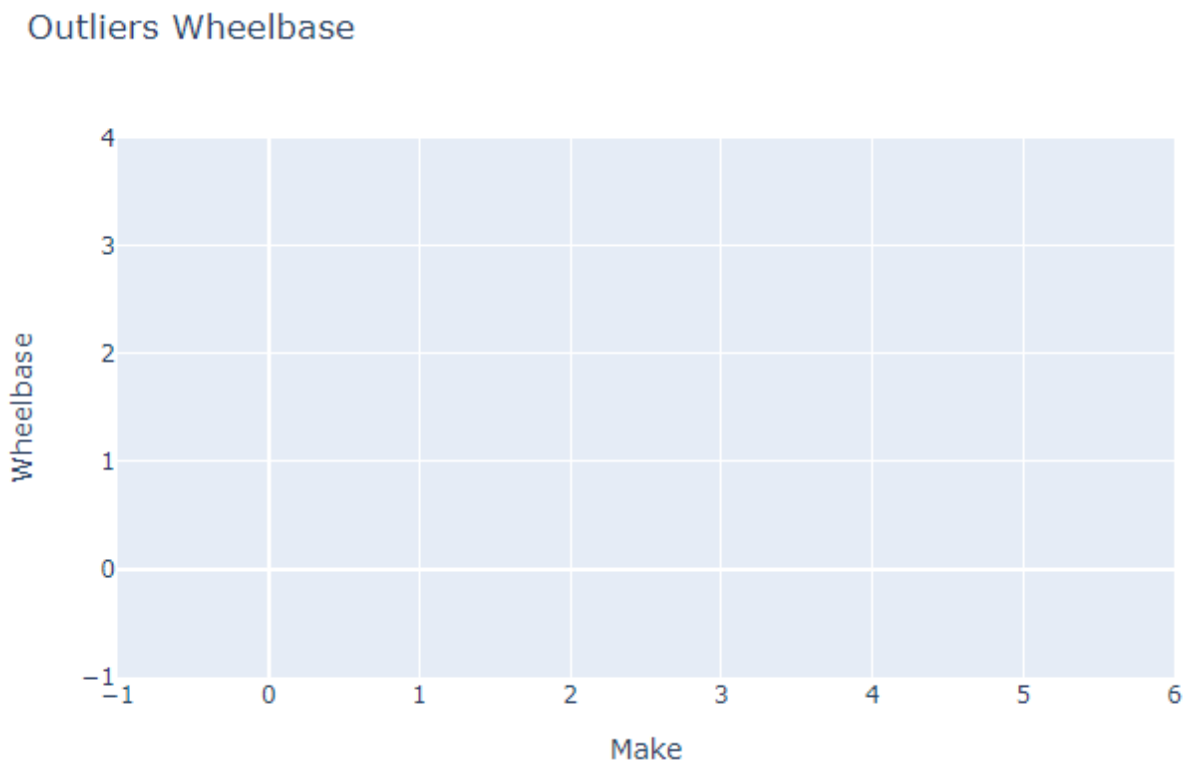

Outliers Cylinders
Cylinders
4
3
2
1
0
−1
−1
0
1
2
3
4
5
6
Make

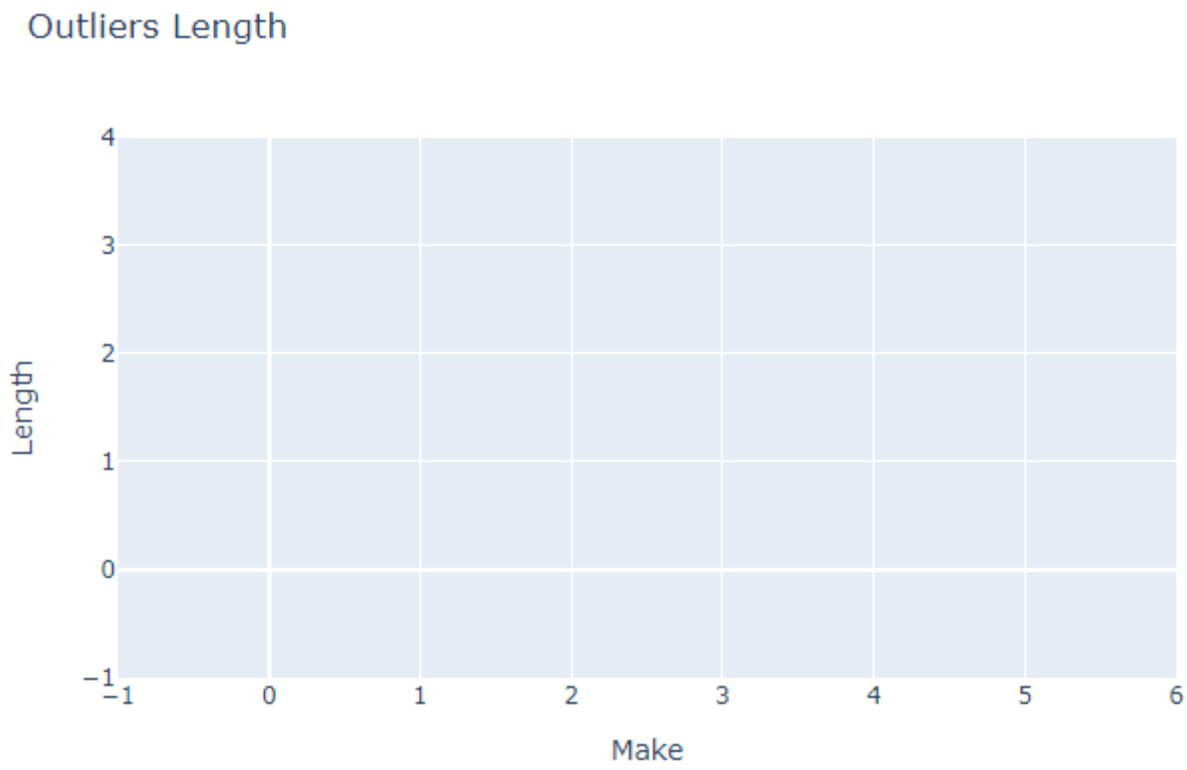

Outliers Horse Power
Horsepower
4
3
2
1
0
−1
−1
0
1
2
3
4
5
6
Make


Outliers Miles per Gallon City
MPG_City
60
59.8
59.6
59.4
59.2
59
Honda
Toyota
Make
Make
Honda
Toyota


Outliers Miles per Gallon Highway
67
66.5
66
65.5
65
MPG_Highway
Honda
Make
Make
Honda


Outliers Weigth
4
3
2
1
0
−1
Weight
−1
0
1
2
3
4
5
6
Make


Outliers Wheelbase
Wheelbase
Make
4
3
2
1
0
−1
−1
0
1
2
3
4
5
6


Outliers Length
Length
Make
4
3
2
1
0
−1
−1
0
1
2
3
4
5
6

3.2. Gráfico de la relación de cada variable con respecto a MPG_City:

3.2.1. Variables categóricas debes crear un boxplot. Explique cómo interpreta el gráfico

3.2.2. Variables numéricas vas a crear un scatter plot.Explique cómo interpreta el gráfico

3.3. Matriz de correlación.

3.3.1. Cree la matriz de correlación, cuáles son las variables más importantes para explicar la variabilidad de MPG_City. Explique por qué el coeficiente es negativo o positivo.

3.3.2. Cree las dummy variables para todas las variables categóricas y genere la matriz de correlación nuevamente. ¿Cuál es el valor de variable categórica con mayor correlación?

3.3.3. Cree la matriz de correlación nuevamente removiendo todos los modelos de carro que fueron catalogados como un outlier. (Puede utilizar. query(‘Model in["MDX","TSX 4dr"]’). Existe alguna variación en la correlación.