Proyecto de Estadística Multivariante Análisis Multivariado

Facultad De Ciencias Naturales y Matemáticas

Dirigido por: Johnny Pambabay Calero, Ph.D.

Oscar Bolaños
(osboflor@espol.edu.ec)

Odalys Tello
(ontello@espol.edu.ec)

Pablo Zúñiga
(pabazuni@espol.edu.ec)

Guayaquil, Noviembre de 2022

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

Como parte del desarrollo y aplicación del contenido aprendido durante el curso de Estadística Multivariante correspondiente a el Segundo periodo académico de 2022 en la Escuela Superior Politécnica del Litroral(ESPOL). Se analíza la base de datos proporcionada por el tutor de curso Johnny Pambabay Calero, Ph.D., la cual contine variables sobre el control de procesos para la produción de alimentos para animales de granja correspondientes al año 2021,se realiza un análisis retrospectivo que ayudara a obtener conclusiones sobre el estado y eventualidades del proceso.

Esta Base de Datos consta de 3823 observaciones cada una de ellas representa a un proceso de elaboracion y sus respectivas mediciones, entre ellas se puede observar a variables cuantitativas como tiempo que demora tomar una orden, temperatura de acondicionamiento entre otras, además de variables categóricas tales como tipo de orden, operador, especie, etc.

En este documento se presenta un reporte sobre el análisis exhaustivo de la Base de Datos bajo Fundamentos Estadísticos permitiendo realizar un análisis descriptivo de los datos, además de Inferencias para finalmente realizar un análisis Multivariante.

Capítulo 2

Objetivos

2.1. Objetivo General

En la busqueda de aprovechar al máximo el conocimiento adquirido durante el desarrollo del curso una vez entendido los fundamentos teóricos se tiene como objetivo general implementar las técnicas aprendidas en un ámbito práctico a los Datos de producción de alimentos para animales mediante las diversas herramientas digitales, bibliograficas poniendo a disposición las habilidades cognitivas, sociales, y sobre todo analíticas que permitan de cierta manera evaluar la comprensión y dominio de los temas vistos además de brindar un aporte a el desarrollo profesional de quienes hacen parte del proyecto.

2.2. Objetivos Específicos

- Realizar un análisis exploratorio de los datos para tomar primeras impresiones respecto a la producción.
- Mediante un análisis descriptivo visualizar las tendencias, frecuencias y comportamiento de las variables para luego ser interpretadas.
- Aplicar las diferentes Técnicas para anaálisis multivariante y lograr una simplificación de los datos que hagan mas sencilla la interpretación de los mismos.

Capítulo 3

Fuente Y Metodología

3.1. Fuente

Los datos obtenidos provienen de una planta de procesamiento la cual produce alimento para animales de granja tales como cerdo, caballos, cuyes, ganado, pavos, aves ponedoras y reproductoras. Este set de Datos cuenta con 41 carácteristicas observadas para cada orden de alimento realizada, en total se tiene el registro de 3823 ordenes.

Tabla 3.1: Datos de los Primeros 6 Procesos

anio	minOrd	NBatch	CapLinea	HdMezcladora	Operador	ESPECIE	TIPO
2021	119	16	17.0	0	ctolagasi	pavos	relacionado
2021	35	1	4.7	0	etoapanta	reproductoras	relacionado
2021	31	1	8.2	0	etoapanta	reproductoras	relacionado
2021	14	1	35.8	0	etoapanta	reproductoras	relacionado
2021	67	7	15.3	0	etoapanta	ponedora rel	relacionado
2021	69	10	18.7	0	etoapanta	cerdos com	comercial

Como se puede ver en la tabla 3.1 la cual contiene 8 características de las 48 que han sido tomadas. Entre las variables observadas podemos encontrar tanto del tipo cualitativas como cuantitativas a continuación se presenta una breve descripción de las mismas:

Cuantitativas

- NBatch : Numero de procesos por lote.
- CapLinea : Capacidad de la Linea
- HMezcla : Húmedad de la mezcla.
- MinOrd : Minutos que lleva tomar una Orden.

Cualitativas

- Operador : Etiqueta de la persona encargada del proceso en curso.

- Tipo : Tipo de proceso y a quien va dirijido, esta cuenta con las categorias relacionado y comercial donde relacionado significa que el proceso esta dirijido a un dominio de la empresa ycomercial se refiere a que la producción esta dirijida una venta particular.
- Especie : Especie a la cual se elabora su producto alimenticio en el proceso.
- Mes : Mes en el cual se esta realizando el proceso.

3.2. Metodología

Durante el Análisis descriptivo es importante realizar en los datos un análisis exploratorio implementando el uso de Histogramas, lo cual permite identificar con que frecuencia los datos toman ciertos valores además medidas de tendencia central tales como la media, mediana, primer y tercer cuartil, nos permiten identificar la ubicacion de los datos y representaciones gráficas como y en base a esto lograr establecer referencias del su comportamiento que serán útiles para realizar inferencias posteriormente.

En el Anáisis Bivariante se obtienen medidas de covariación, correlaciones entre otras con las cuales podemos identificar como nuestars variables interactuan entre ellas permitiendo realizar modelaciones Mediante tecnicas como Regresión multiple, uncluso se pueden representar ciertos grupos de variables mediante un diagramas de cajas para comparar su eficiencia en los procesos.

Para el análisis Multivariante se implementa técnicas como Análisis de Componentes Principales (ACP), Análisis Factorial(AF), Análisis de Correlación Canónico(ACC), Escalamiento Multidimensiona(EM), Análisis de Conglomerados(AC) y Análisis Discriminante(AD), estas técnicas antes mencionadas nos permiten realizar un análisis de los datos mediante formuaciones matemáticas que nos ayudan a lograr una mejor interpretación de los datos que no serían faciles de lograr sin estas Técnicas.

Capítulo 4

Análisis Descriptivo de Datos

Un analisis descriptivo y gráfico de las variables es de gran utilidad para identificar la naturaleza de las misamas perimitiendo una mejor interpretacion de estas y a su vez inferencias mas relevantes, a continuación se presenta el análisis antes mencionado para variables tanto cualitativas como cualitativas.

4.1. Variables Aleatorias Cualitativas

4.1.1. Variable Operador

En la 4.1, se describe la frecuencia de órdenes de venta que realizó cada operador en donde el menor productor fue pquishpe con solamente 2 producciones, pero esto es atribuido a que los meses de producción son menos que el resto de operadores, solo hay producción en el mes de mayo y agosto. Por otra parte, el operador con mayor producción fue ccondor con 1299 producciones.

Tabla 4.1: Tabla de Frecuencia del Operador

Operador	Frecuencia
ccondor	1299
ctolagasi	1110
etoapanta	12
jbone	44
jcarguacundo	232
kfarinango	76
pquishpe	2
wpazmino	1048

El gráfico 4.1 sepuede apreciar de manera gráfica y colorida, quizás este se puede entender un poco mejor para quienes es un más difícil entender la tabla de frecuencias.

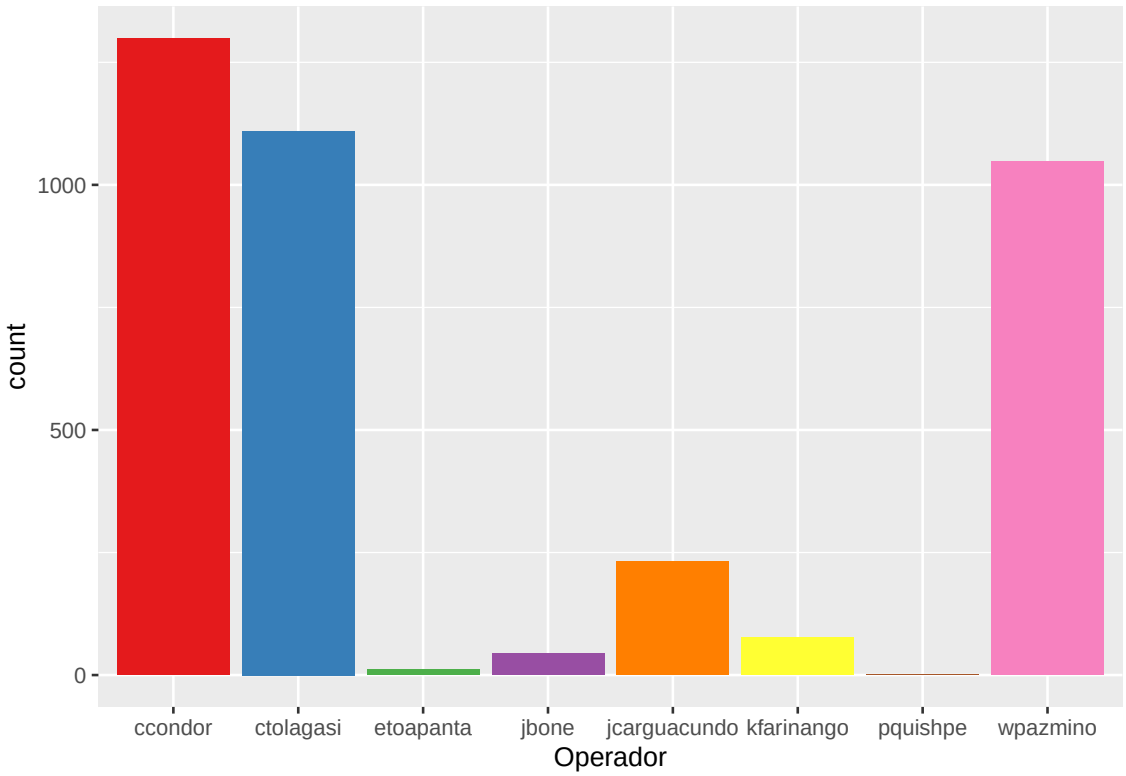

Figura 4.1: Histograma Operador.

4.1.2. Variable Tipo

En esta tabla 4.2, se habla del número total de órdenes que fuero para destino comercial y relacionado, entonces se observa que la mayor cantidad de ordenes generadas son para la venta comercial.

Tabla 4.2: Tabla de Frecuencia del Tipo

Tipo	Frecuencia
comercial	2754
relacionado	1069

En el gráfico 4.2 se aprecia el gráfico de frecuencias que hace referencia a la tabla 4.2.

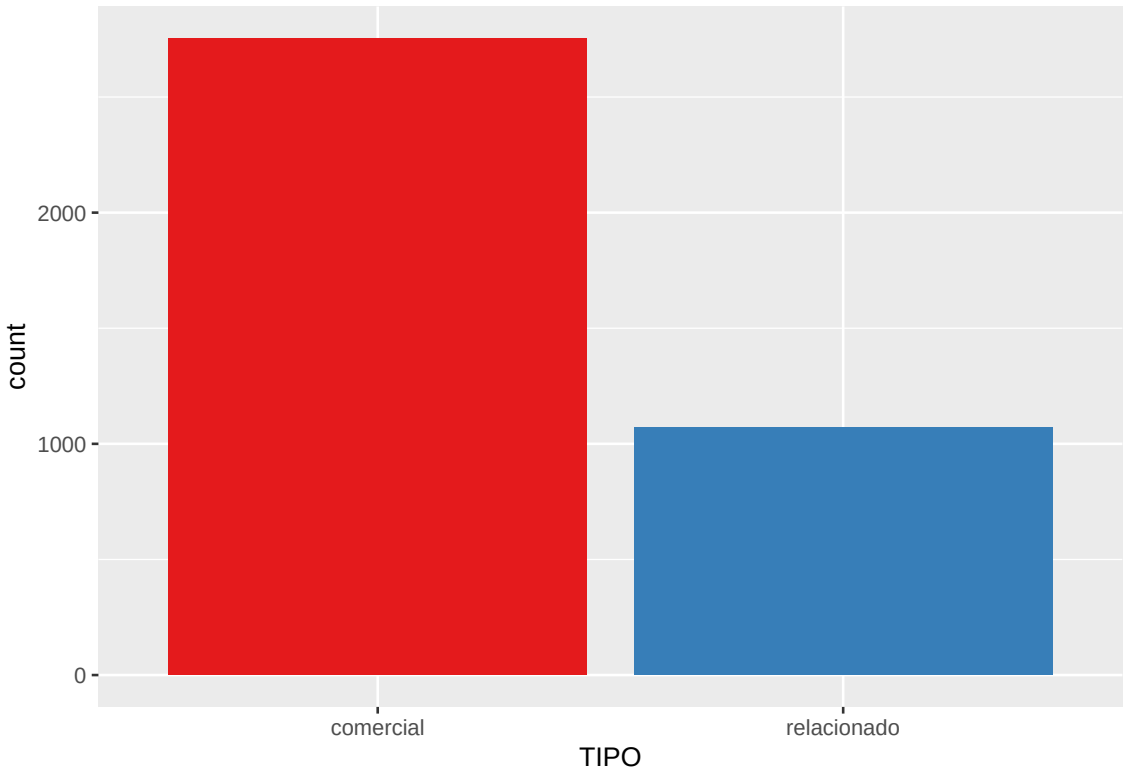

Figura 4.2: Histograma Tipo.

4.1.3. Variable Especie

En la Tabla 4.3, menciona la cantidad de balanceado producido para cada especie y la que más se produjo en todo el año fue para cerdos y para venta comercial y el menor fue para ponedora comercial.

Tabla 4.3: Tabla de Frecuencia de la Especie

Especie	Frecuencia
caballos	175
cerdos com	1508
cerdos rel	364
cuyes y conejos	104
engorde com	264
engorde rel	33
ganado	694
pavos	339
ponedora com	9
ponedora rel	131
reproductoras	202

La figura 4.3 describe aquello que menciona la tabla 4.3 de manera gráfica entonces se puede apreciar lo que se escribió antes.

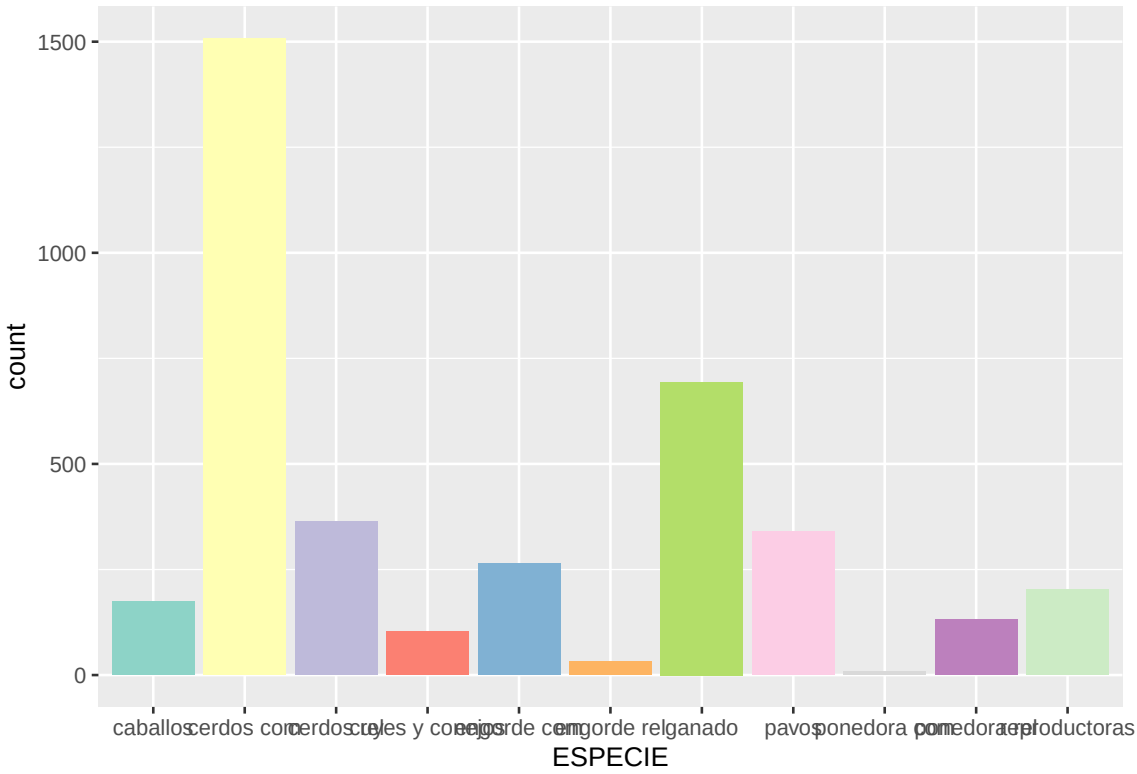

Figura 4.3: Histograma Especie.

4.1.4. Variable Mes

En la siguiente tabla 4.4 se describe la producción de cada mes en donde el mes con mejor producción es el mes de octubre y el de mejor producción es el mes de febrero.

Tabla 4.4: Tabla de Frecuencia del Mes

Mes	Frecuencia
Enero	315
Febrero	266
Marzo	309
Abril	308
Mayo	336
Junio	296
Julio	347
Agosto	319
Septiembre	339
Octubre	351
Noviembre	307
Diciembre	330

La imagen 4.4 muestra de manera gráfica lo que describe la tabla 4.4.

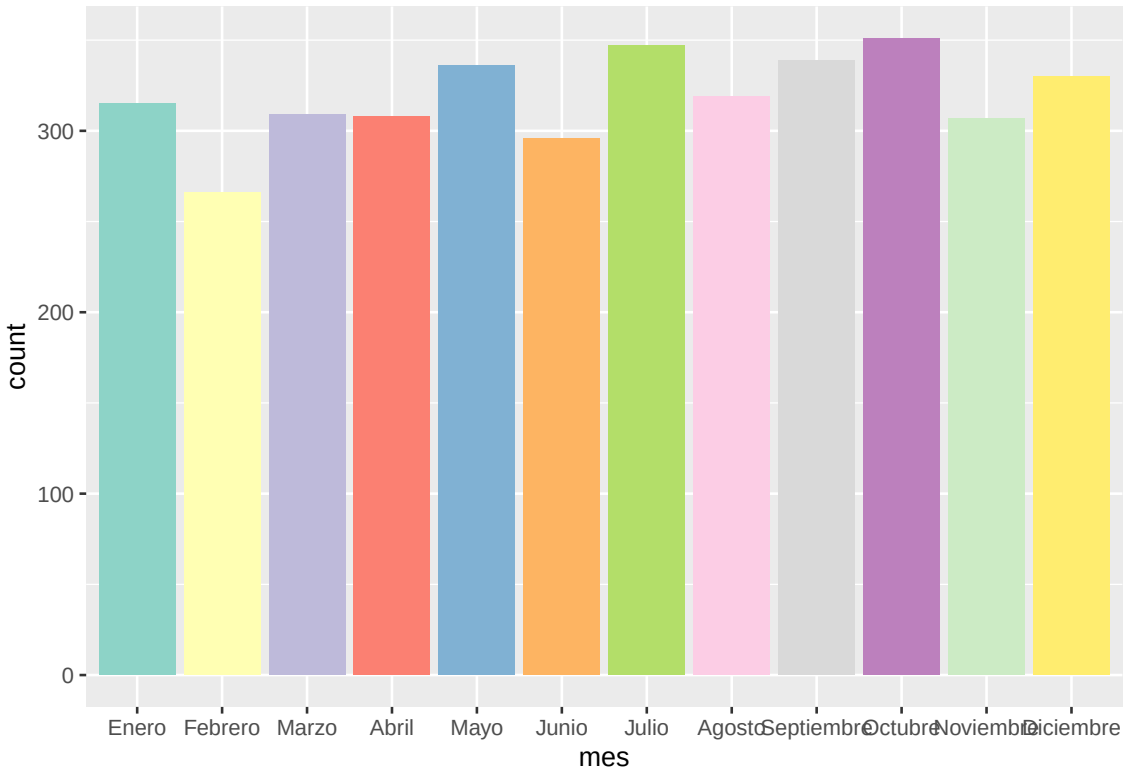

Figura 4.4: Histograma Mes.

4.2. Variables Aleatorias Cuantitativas

4.2.1. Variable NBATCH

Esta índica el numero de procesos que se han realizado de un solo lote de materia prima, podemos notar que con una mayor frecuencia(ver tabla 4.5), especificamente el 48.37 % de numero de procesos por lote se encuentra entre 0 y 8 seguido con un 21.11 % entre 8 y 17 procesos por lote.

Tabla 4.5: Tabla de Frecuencia del NBatch

Class limits	f	rf	rf(%)	cf	cf(%)
[0,8.5462)	1849	0,48	48,37	1849	48,37
[8.5462,17.092)	807	0,21	21,11	2656	69,47
[17.092,25.638)	626	0,16	16,37	3282	85,85
[25.638,34.185)	274	0,07	7,17	3556	93,02
[34.185,42.731)	110	0,03	2,88	3666	95,89
[42.731,51.277)	120	0,03	3,14	3786	99,03
[51.277,59.823)	7	0,00	0,18	3793	99,22
[59.823,68.369)	20	0,01	0,52	3813	99,74
[68.369,76.915)	3	0,00	0,08	3816	99,82
[76.915,85.462)	4	0,00	0,10	3820	99,92
[85.462,94.008)	1	0,00	0,03	3821	99,95
[94.008,102.55)	1	0,00	0,03	3822	99,97
[102.55,111.1)	1	0,00	0,03	3823	100,00

	x
start	0,00
end	111,10
h	8,55
right	0,00

Estos porcentajes se pueden visualizar mas claramente en su respectivo histograma(ver en la figura 4.5) es evidente que numeros altos de procesos por lote sean menos frecuentes esto puede deberse a tipos de producción especiales.

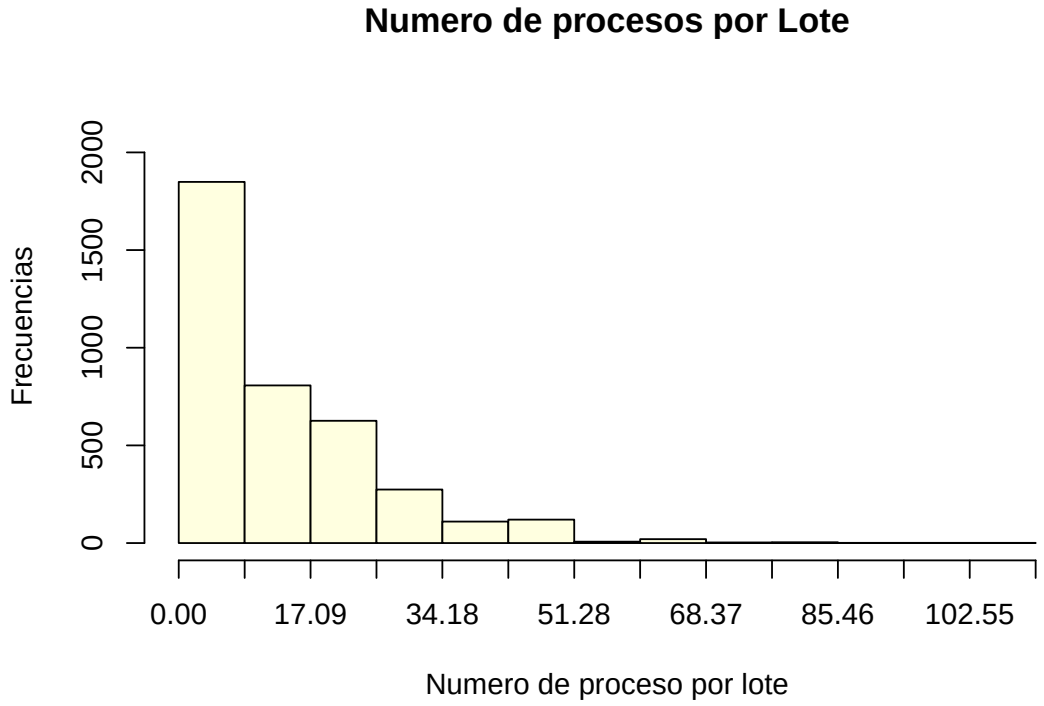

Figura 4.5: Histograma NBATCH.

Entre las medidas de tendencia central(figura4.6) se pueden sacar conclusiones interesantes como que el 75 % de el numero de procesos por lote se encuentra por debajo de

20, adicionalmete la media de los datos esta en 13.68 mientras que la mediana es 10 con lo cual podemos identificar un sesgo relativo.

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00    5.00   10.00   13.68   20.00  110.00
```

Figura 4.6: MTC NBATCH.

4.2.2. Variable CAPLINEA

Esta variable mide la capacidad de la Linea en la cual se esta trabajando, se puede notar que con una frecuencia de 99.22 % esta se encuentra entre 0 y 140 es decir la primera clase(ver talbla 4.6), bajo la métrica en la que se mide dicha capacidad.

Tabla 4.6: Tabla de Frecuencia del CapLinea

Class limits	f	rf	rf(%)	cf	cf(%)
[0,139.7296)	3793	0,99	99,22	3793	99,22
[139.7296,279.4592)	20	0,01	0,52	3813	99,74
[279.4592,419.1888)	3	0,00	0,08	3816	99,82
[419.1888,558.9185)	4	0,00	0,10	3820	99,92
[558.9185,698.6481)	0	0,00	0,00	3820	99,92
[698.6481,838.3777)	0	0,00	0,00	3820	99,92
[838.3777,978.1073)	0	0,00	0,00	3820	99,92
[978.1073,1117.837)	2	0,00	0,05	3822	99,97
[1117.837,1257.567)	0	0,00	0,00	3822	99,97
[1257.567,1397.296)	0	0,00	0,00	3822	99,97
[1397.296,1537.026)	0	0,00	0,00	3822	99,97
[1537.026,1676.755)	0	0,00	0,00	3822	99,97
[1676.755,1816.485)	1	0,00	0,03	3823	100,00

	x
start	0,00
end	1816,48
h	139,73
right	0,00

Esto es más visible en el respectivo histograma(figura 4.7), se puede interpretar que esta alta frecuencia se deba a que se espera una capacidad stándar para el proceso y que esta se encuentre en esta primera clase, sin embargo los que estan fuera de la primera clase puede deberse a un tipo de producción especial para algun producto específico o simplemente un fuera de control en el proceso.

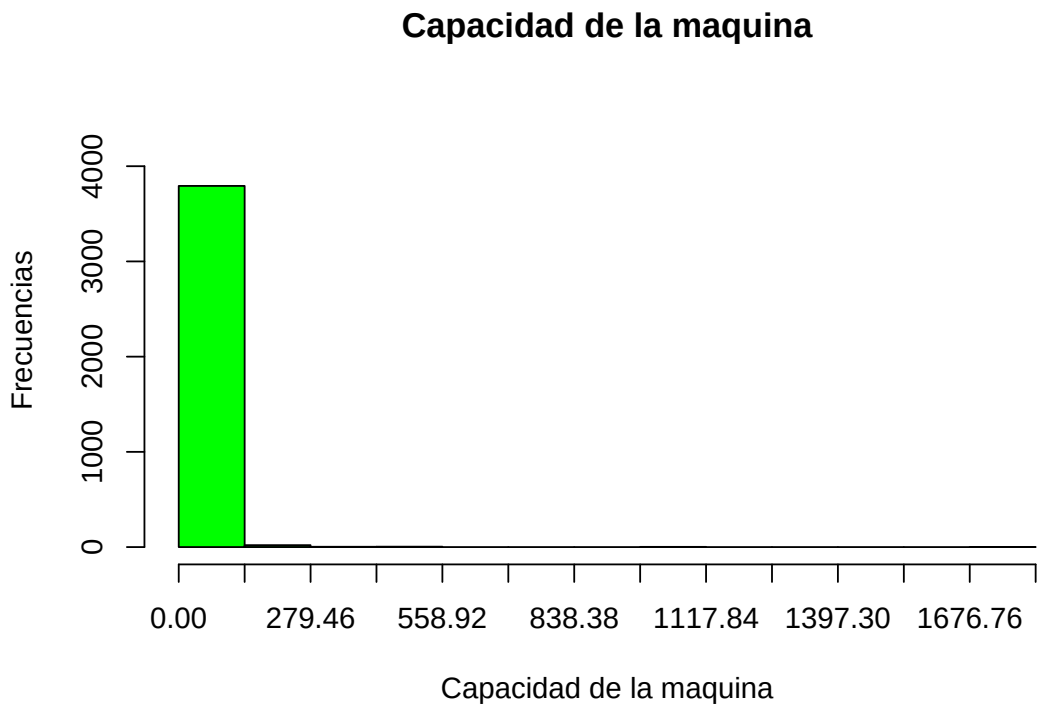

Figura 4.7: Histograma CAPLINEA.

Se puede interpetar que en el 75 % de los procesos la capacidad de la linea esta por debajo de 22.70 lo cual es cercano a la media la cual es 23.12 se podría decir que este es el valor stándar anteriormente mencioado, también podria verse como la capácidad máxima si se pensara en que todas deben funciaonar a máxima capacidad en terminos de óptimizacion de la producción.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00	16.30	19.60	23.12	22.70	1798.50

Figura 4.8: MTC CAPLINEA.

4.2.3. Variable HMEZCHLA

Esta variable representa la humedad de la mezcla durante cierta etapa de la producción esta puede ser considerarada una variable importante, pues un descontrol de la misma podria afectar a la linea de prodcucción hechando a perder materia prima, podemos evidenciar una alta presencia de humedad entre 7.77 y 15.5 en los procesos, y una presencia considerable en 0 y 7.77 con un porcentaje de 79.49 % y 20.09 % respectivamente (ver tabla 4.7).

Tabla 4.7: Tabla de Frecuencia del Hmezcla

Class limits	f	rf	rf(%)	cf	cf(%)
[0,7.77)	768	0,20	20,09	768	20,09
[7.77,15.5)	3039	0,79	79,49	3807	99,58
[15.5,23.3)	7	0,00	0,18	3814	99,76
[23.3,31.1)	3	0,00	0,08	3817	99,84
[31.1,38.8)	1	0,00	0,03	3818	99,87
[38.8,46.6)	0	0,00	0,00	3818	99,87
[46.6,54.4)	0	0,00	0,00	3818	99,87
[54.4,62.2)	3	0,00	0,08	3821	99,95
[62.2,69.9)	0	0,00	0,00	3821	99,95
[69.9,77.7)	0	0,00	0,00	3821	99,95
[77.7,85.5)	1	0,00	0,03	3822	99,97
[85.5,93.2)	0	0,00	0,00	3822	99,97
[93.2,101)	1	0,00	0,03	3823	100,00

	x
start	0,00
end	101,00
h	7,77
right	0,00

Se desea que la temepratura este controlada y sea especifica, esta puede diferir dependiendo del tipo de producto que se este procesando en el momento, podemos atribuirle esto a que se presenten altas frecuencias en las clases mencionadas anteriormente, asignandole a cada clase un producto especifico que requiere una humedad también especifica, esto es mas sencillo de ver en el histograma que se muestra en la figura 4.9.

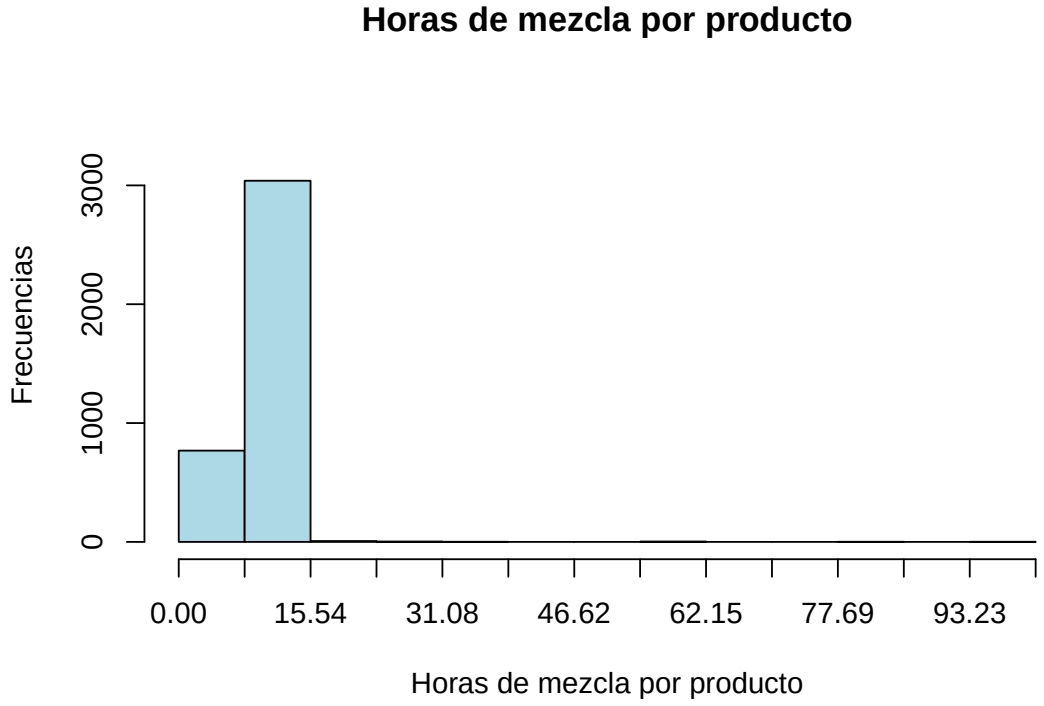

Figura 4.9: Histograma CAPLINEA.

Podemos notar que el 75 % de las humedades registradas estan por encima de 8.27

ademas que la temperatura media se encuentra en 7.68 y que se ha resgsitrado una temperatura máxima de hasta 100 unidades en base a la metrica seleccionada para medir la Humedad de la mezclas.

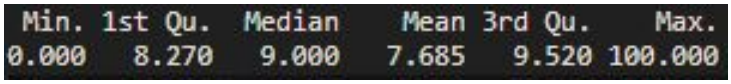

Figura 4.10: MTC HMEZCLA.

4.2.4. Variable MINORD

Esta variable cualitativa mide en minutos el tiempo que toma procesar una orden desde el momenteo en que se toma la orden hasta que esta esta lista para ser entregada. Para esta variable presenta una alta frecuencia en la clase que se encuentra desde 0 hasta 149.40 con un 84.07 % de presencia tambien existe una frecuencia considerable en la clase que va de 149.40 a 289.80.

Tabla 4.8: Tabla de Frecuencia del minOrd

Class limits	f	rf	rf(%)	cf	cf(%)
[0,149.4023)	3214	0,84	84,07	3214	84,07
[149.4023,298.8046)	511	0,13	13,37	3725	97,44
[298.8046,448.2069)	82	0,02	2,14	3807	99,58
[448.2069,597.6092)	9	0,00	0,24	3816	99,82
[597.6092,747.0115)	3	0,00	0,08	3819	99,90
[747.0115,896.4138)	1	0,00	0,03	3820	99,92
[896.4138,1045.816)	0	0,00	0,00	3820	99,92
[1045.816,1195.218)	2	0,00	0,05	3822	99,97
[1195.218,1344.621)	0	0,00	0,00	3822	99,97
[1344.621,1494.023)	0	0,00	0,00	3822	99,97
[1494.023,1643.425)	0	0,00	0,00	3822	99,97
[1643.425,1792.828)	0	0,00	0,00	3822	99,97
[1792.828,1942.23)	1	0,00	0,03	3823	100,00

	x
start	0,00
end	1942,23
h	149,40
right	0,00

En el Histograma podemos notar como altos tiempos de procesamiento por cada orden son menos frecuentes(ver figura 4.11), nuevamente la óptimizacion de la producción juega un papel clave en esta interpretación pues un tiempo de produccíon óptimo seria lo adecuado.

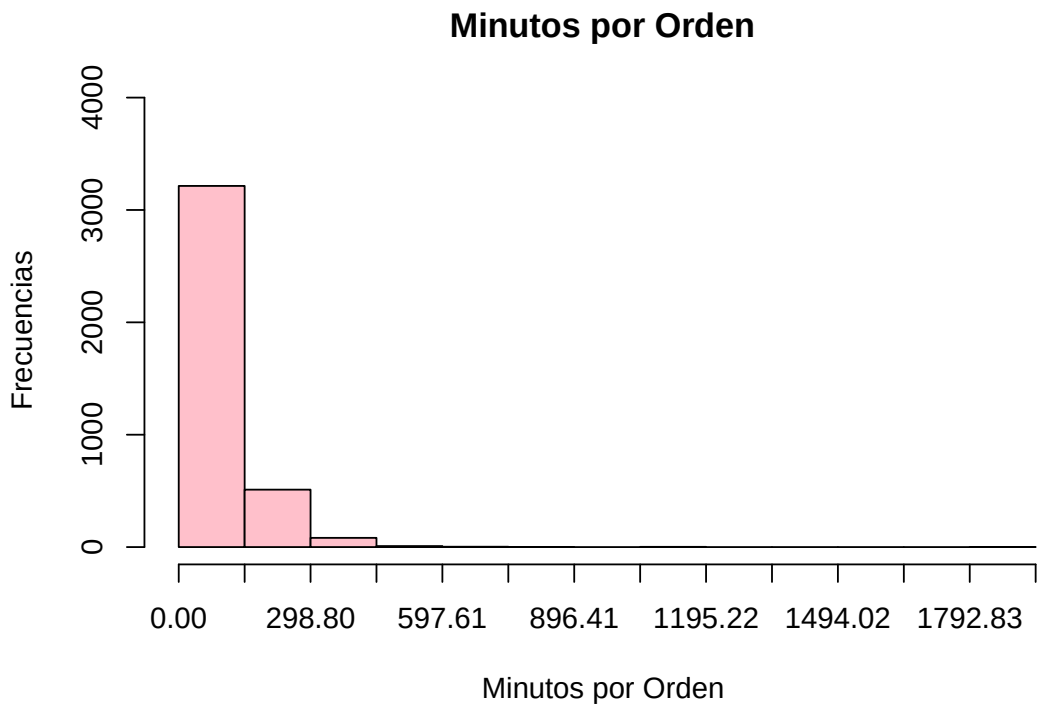

Figura 4.11: Histograma MINORD.

Podemos ver entonces en la figura 4.12 que la media esta en arpoximadamente 87 minutos mientras que su mediana es 60 minutos, es importante notar que el tiempo que toma procesar una orden debe estar altamente correlacionado con el tipo de producto que se esta produciendo siendo así que el "producto estrella" sea aquel que tome aproximadamente 87 minutos.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00	31.00	60.00	86.97	115.00	1923.00

Figura 4.12: MTC MINORD.

Capítulo 5

Análisis Estadístico Bivariante

5.1. Matríz de Correlación

En la figura 5.1 se muestra la tabla de correlaciones entre las siguientes variables: temperatura de acondicionamiento inicial, temperatura de acondicionamiento final, velocidad del alimentador y la temperatura del expandido y, se muestra una alta correlación (definiendo como alta correlación arriba del 0.6) entre la temperatura de acondicionamiento inicial y final.

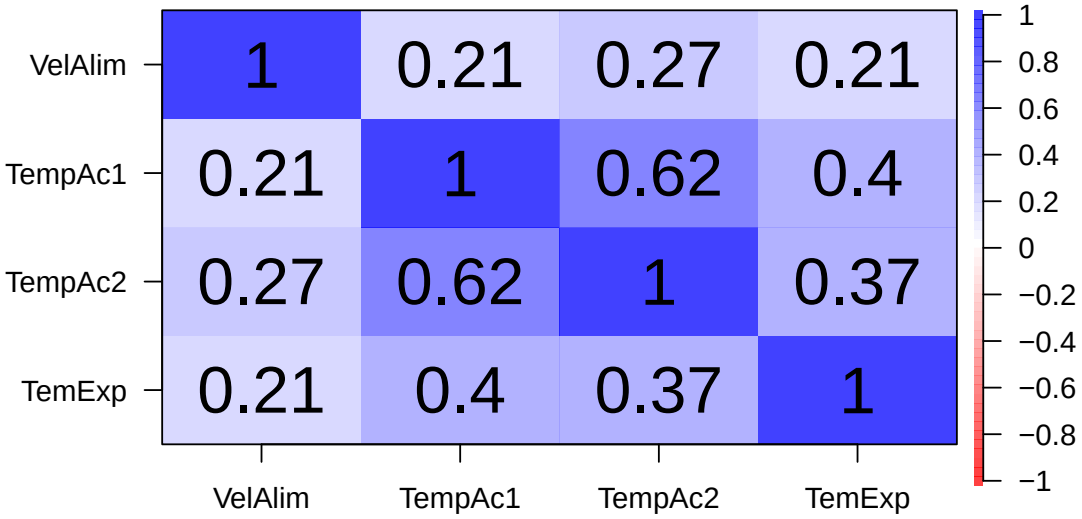

Figura 5.1: Correalciones.

5.2. Varianzas y Covarianzas

La tabla 5.1 muestra la varianza y covarianza entre las variables: temperatura de acondicionamiento inicial, temperatura de acondicionamiento final y la temperatura del expandido.

Tabla 5.1: Varianzas y Covarianzas

	Temp. Ac1	Temp. Ac2	Temp Exp
TempAc1	67,9	44,4	67,5
TempAc2	44,4	75,6	65,5
TemExp	67,5	65,5	408,9

5.3. Diagrama de Cajas

En la figura 5.2 se muestran los diagramas de cajas de las 3 variables antes mencionadas, lo que se puede apreciar es que en temperatura de acondicionamiento 2 y temperatura de expandido hay varios valores atípicos.

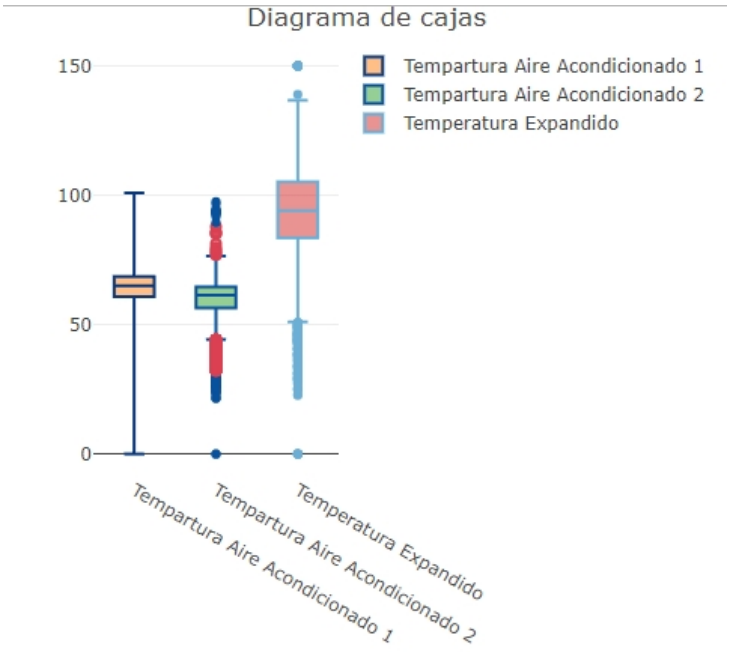

Figura 5.2: Diagrama de Cajas

5.4. Modelo de Regresión

En la figura 5.3 se muestra un modelo de regresión lineal que trata de modelar la Temperatura de acondicionamiento inicial vas la temperatura de acondicionamiento final.

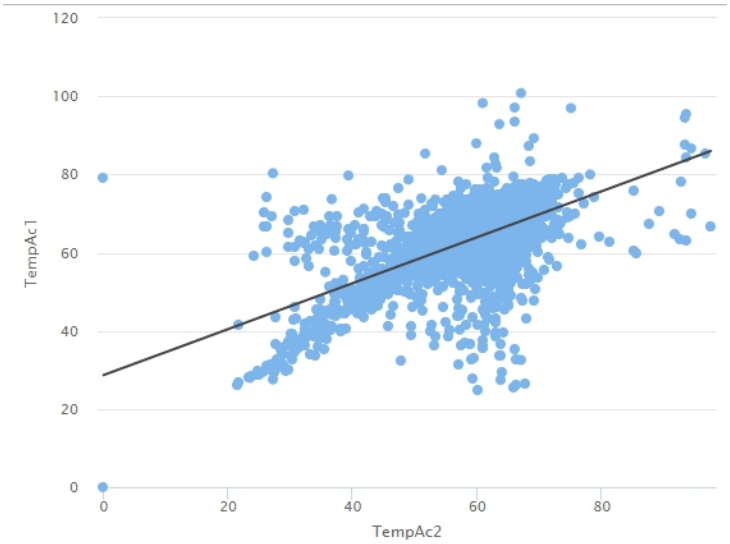

Figura 5.3: Regresión Lineal

Capítulo 6

Estadística Inferencial

6.1. Bondad de ajuste

En la figura 6.1 se muestra una prueba de bondad de ajuste donde la hipótesis nula plantea que la media de la temperatura de acondicionamiento inicial está situada en menos de 50 grados y la hipótesis alterna trata de explicar lo contrario. Con un valor p de 0.001948 se rechaza la hipótesis nula.

SUPUESTOS	Estadísticas Descriptivas					
Muestras grandes, el tamaño de la muestra n es mayor o igual que 30	Media	Error estándar de la media	Desviación estandar	Sesgo	Mediana	Moda
	63.604	0.133	8.239	-1.435	64.99	66.97
Contraste de Hipótesis	Estadístico de prueba				Valor p	
H0: μ < 50 vs **H1:** μ >= 50	$\chi^2 = \sum \frac{(O_i - e_i)^2}{e_i}$		χ^2= 4079.3		P-value= 0.001948	
	Conclusión: Ya que el valor de p = 0.001948, rechazamos la hipótesis nula que nos dice que la media de la temperatura de acondicionamiento 1 es menor a 50.					

Figura 6.1: Bondad de ajuste TEMPAC1

6.2. Pruebas de Hipótesis e Intervalos de confianza

6.2.1. TEMP_AC1

La figura 6.2 muestra un contraste de hipótesis donde la hipótesis nula plantea exactamente lo mismo que en la figura 6.1 pero aquí se plantea un intervalo de confianza del 95 % que dice que la temperatura de acondicionamiento inicial va desde los 50 grados hasta los 77.20 grados.

SUPUESTOS	Estadísticas Descriptivas					
Muestras grandes, el tamaño de la muestra n es mayor o igual que 30.	Media	Error estándar de la media	Desviación estandar	Sesgo	Mediana	Moda
	63.604	0.133	8.239	-1.435	64.99	66.97
Contraste de Hipótesis	Estadístico de prueba				Valor p	
	$Z=\frac{\bar{X}-\mu}{s/\sqrt{n}}$		Z= 102.0909		P-value= 0.001948	
	Intervalo de confianza					
H0: μ < 50 vs **H1:** μ >= 50	$\bar{X}-z_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{X}+z_{\alpha/2}\frac{s}{\sqrt{n}}$		50.000 < μ < 77.208			
Conclusión: Con un intervalo de confianza del 95% podemos asegurar que la media de la temperatura de acondicionamiento 1 se va a encontrar entre [50.00000, 77.20841].						

Figura 6.2: P. Hipoótesis e IC TEMPAC1

6.2.2. TEMP_AC2

La figura 6.3 muestra un contraste de hipótesis donde la hipótesis nula plantea exactamente lo mismo que en la figura 6.1 pero aquí se plantea un intervalo de confianza del 95 % que dice que la temperatura de acondicionamiento final va desde los 50 grados hasta los 68.79 grados.

SUPUESTOS	Estadísticas Descriptivas					
Muestras grandes, el tamaño de la muestra n es mayor o igual que 30.	Media	Error estándar de la media	Desviación estandar	Sesgo	Mediana	Moda
	59.399	0.140	8.692	-1.361	61.35	65.02
Contraste de Hipótesis	Estadístico de prueba				Valor p	
	$Z=\frac{\bar{X}-\mu}{s/\sqrt{n}}$		Z= 66.856		P-value= 2.2e-16	
	Intervalo de confianza					
H0: μ < 50 vs **H1:** μ >= 50	$\bar{X}-z_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{X}+z_{\alpha/2}\frac{s}{\sqrt{n}}$		50.000 < μ < 68.79845			
Conclusión: Con un intervalo de confianza del 95% podemos asegurar que la media de la temperatura de acondicionamiento 2 se va a encontrar entre [50.00000, 68.79845].						

Figura 6.3: P. Hipoótesis e IC TEMPAC2

6.3. Análisis de Contingencia y Pruebas de Hipótesis

6.3.1. Tipo y Especie

Se realiza una tabla de contingencia para las variables Tipo y Especie, ubicando a esta ultima en las columnas y a Tipo en las filas como se puede ver en la figura 6.4.

	Caballo	Cerdos com	Cerdos rel	Cuyes y conejos	Engorde com	Engorde rel	Ganado	Pavos	Ponedora com	Ponedora rel	Reproductoras
Comercial	175	1508	0	104	264	0	694	0	9	0	0
Relacionado	0	0	364	0	0	33	0	339	0	131	202

Figura 6.4: P. Tabla de Contingencia para TIPO y ESPECIE

Se puede notar que en terminos de produccion en base a la especies a quien va dirigida la producción del tipo relacionado no es del todo mayor a la del tipo comercial, para constatar se plantea una prueba de Hipótesis con el estadístico Chi-Cuadrado(ver figura 6.5).

Contraste de Hipótesis	Estadístico de prueba		Valor p
H0: La cantidad de especies de tipo relacionado es mayor al de tipo comercial **H1:** La cantidad de especies de tipo relacionado NO mayor al de tipo comercial	$\chi^2 = \sum \frac{(O_i - e_i)^2}{e_i}$	χ^2= 3823	P-value= 2.2e-16
Conclusión: Ya que el valor de p = 2.2e-16, rechazamos la H0 que nos dice que La cantidad de especies de tipo relacional es mayor al de tipo comercial.			

Figura 6.5: P. Prueba de Hipótesis

Se obtiene un valor p de 2.2e-16 con lo cual se rechaza la hipotesis nula por lo tanto La cantidad de especies del tipo relacionado No es mayor a la del tipo comercial.

6.3.2. Tipo y Operador

Se realiza una tabla de contingencia para las variables Tipo y Operador, similar a como se hizo con la variable especie se asigna a Operador para las columnas y Tipo para las filas(figura 6.6).

	Ccondor	Ctologasi	Etoapanta	Jbone	Jcarguacundo	Kfarinango	Pquishpe	Wpazmino
Comercial	950	824	6	33	167	56	2	716
Relacionado	349	286	6	11	65	20	0	332

Figura 6.6: P. Tabla de Contingencia para TIPO y OPERADOR

Se puede de igual manera notar que en terminos de produccion en base a la producción de los operadores los procesos dirigidos al tipo relacionado no es del todo mayor a la del tipo comercial,verificar esto se plantea una prueba de Hipótesis con el estadístico Chi-Cuadrado(ver figura 6.7).

Contraste de Hipótesis	Estadístico de prueba		Valor p
H0: La cantidad de ventas de tipo relacional con respecto a los operadores es mayor al de tipo comercial **H1:** La cantidad de ventas de tipo relacional con respecto a los operadores NO es mayor al de tipo comercial	$\chi^2 = \sum \frac{(O_i - e_i)^2}{e_i}$	χ^2= 14.585	P-value= 0.0417
Conclusión: Ya que el valor de p = 0.0417, rechazamos la H0 que nos dice que La cantidad de ventas de tipo relacional con respecto a los operadores es mayor al de tipo comercial			

Figura 6.7: P. Prueba de Hipótesis

Se obtiene un valor p de 0.0417 con lo cual se rechaza la hipotesis nula por lo tanto La cantidad de producciones del tipo relacionado No es mayor a la del tipo comercial.

Capítulo 7

Estadística Multivariante

7.1. Análisis de Correlación Canónico

En la figura 7.1 se muestra la matriz de correlación entre las siguientes variables: NBatch, VelAlim, TempAc1, TempAc2, TempExp, PresionCono, minOrd, CapLinea, CantKgProd y Granulometria. ### Matriz de Correlación de los Datos

Tabla 7.1: Matriz de Correlaciones

	NBatch	VelAlim	TempAc1	TempAc2	TemExp	PresionCono	minOrd	CapLinea	CantKgProd	Granulometria
X1	1.000	0.457	0.222	0.293	0.262	-0.093	0.720	0.028	0.994	-0.161
X2	0.457	1.000	0.210	0.269	0.206	0.079	0.344	-0.083	0.448	-0.085
X3	0.222	0.210	1.000	0.620	0.405	0.083	0.082	-0.016	0.222	-0.005
X4	0.293	0.269	0.620	1.000	0.373	0.069	0.139	-0.058	0.291	-0.059
X5	0.262	0.206	0.405	0.373	1.000	0.000	0.195	-0.018	0.261	-0.036
X6	-0.093	0.079	0.083	0.069	0.000	1.000	-0.115	-0.041	-0.083	0.252
Y1	0.720	0.344	0.082	0.139	0.195	-0.115	1.000	-0.059	0.713	-0.139
Y2	0.028	-0.083	-0.016	-0.058	-0.018	-0.041	-0.059	1.000	0.034	-0.088
Y3	0.994	0.448	0.222	0.291	0.261	-0.083	0.713	0.034	1.000	-0.159
Y4	-0.161	-0.085	-0.005	-0.059	-0.036	0.252	-0.139	-0.088	-0.159	1.000

7.1.1. Correlación entre las variables Canónicas

Se presenta la correlación entre las variables canónicas:

Tabla 7.2: Correlación entre las variables Canónicas

	V1	V2	V3	V4
U1	0.989	0.00	0.000	0.000
U2	0.000	0.07	0.000	0.000
U3	0.000	0.00	0.022	0.000
U4	0.000	0.00	0.000	0.014

7.1.2. Combinaciones lineales

A contiunuación se muestra que de las ecuaciones se obtienen los coeficientes de las combinaciones lineales y así, las variables canónicas asociadas resultantes.

$$\begin{cases} U_1 = a^{(1)'}X = 7.928e^{-02}X_{NBatch} - 4.668e^{-04}X_{VelAlim} + 2.292e^{-04}X_{TempAc1} - 4.084e^{-04}X_{TempAc2} \\ +9.184e^{-05}X_{TempExp} + 1.345e^{-03}X_{PresionCono} \\ V_1 = b^{(1)'}Y = 2.223e^{-04}Y_{minOrd} - 8.633e^{-05}Y_{CapLinea} + 4.599e^{-05}Y_{CantKgProd} - 5.859e^{-08}Y_{Granulometria} \end{cases}$$

$$\begin{cases} U_2 = a^{(1)'}X = -0.014X_{NBatch} + 0.013X_{VelAlim} - 0.034X_{TempAc1} + 0.012X_{TempAc2} \\ +0.002X_{TempExp} - 0.133X_{PresionCono} \\ V_2 = b^{(1)'}Y = 5.782e^{-03}Y_{minOrd} + 8.305e^{-05}Y_{CapLinea} - 2.410e^{-05}Y_{CantKgProd} - 1.823e^{-03}Y_{Granulometria} \end{cases}$$

7.2. Análisis de Componentes Principales

La figura 7.1 muestra de los puntos proyectados sobre el eje Y1 y Y2 se puede vver que las temperaturas de acondicionamiento estan relacionadas con la primera componente mientras que la presion del cronometro esta relacionada con la segunda componente.

7.2.1. Bitplot sin Rotación

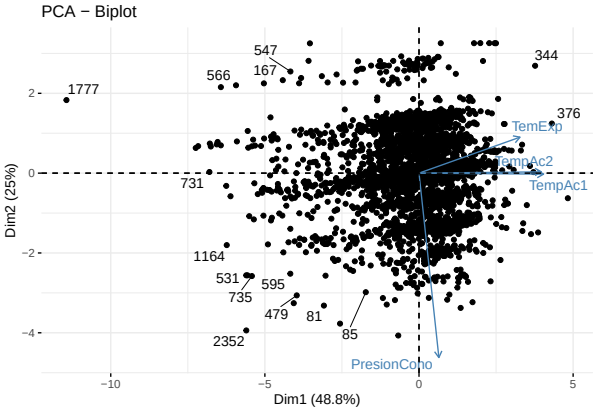

Figura 7.1: Bitplot sin Rotación.

7.2.2. Bitplot con Rotación varimax

En la figura 7.2 se ve un cambio en la relacion ahora positiva con PresionCrono y la segunda componente principal, ademas las Temperaturas de acondicionamiento se alejan minimamente de la primera componente principal.

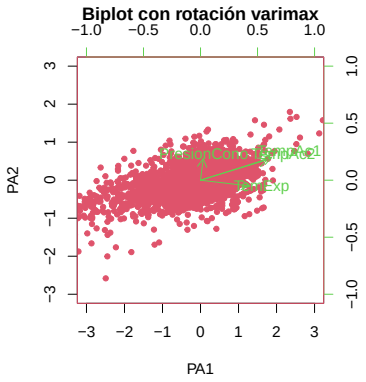

Figura 7.2: Bitplot sin Rotación.

7.2.3. Porcentaje de Varianza Explicada

Se puede ver en la figura 7.3 que el mayor porcentaje de la variabilidad de los datos esta inmersa en las 3 primeras componentes abarcando un 90.6 % de variabilidad.

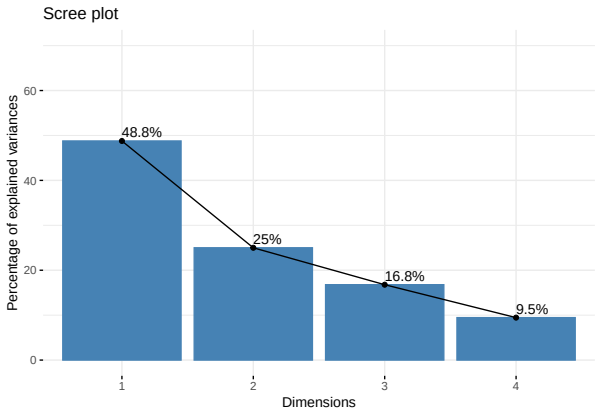

Figura 7.3: Porcentaje de Varianza Explicada.

7.2.4. Gráfico de Correlación

A continuacion se mmuestran las correalaciones entre las cuales se puede ver que la correlacion es alta entre Temp_ac1 y Temp_ac2.

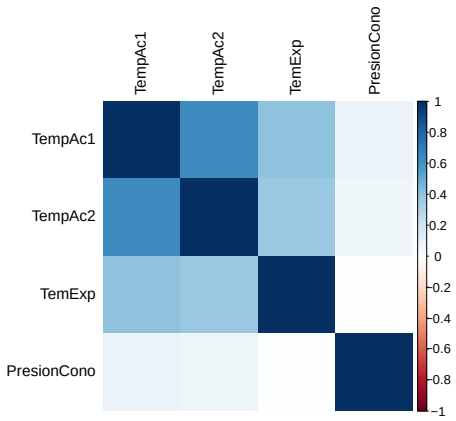

Figura 7.4: Gráfico de Correlación.

7.3. Análisis Factorial

7.3.1. Bitplot sin rotacion

Se puede ver obviando las comparaciones con respecto a ACP, en la figura 7.5 que existe una mayor correlación de las temperaturas de acondicinamiento y expandido con el primer factor común, mientras que la presión del crono esta mas relacionado a el segundo factor común.

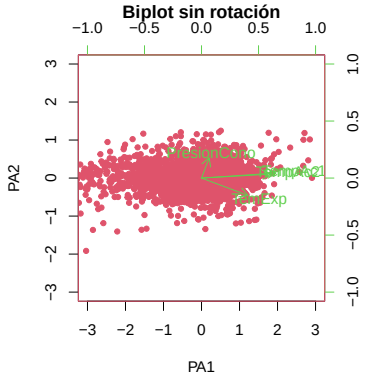

Figura 7.5: Bitplot sin Rotación af.

7.3.2. Bitplot con rotación quartimax

Aplicando una rotación con el objetivo de tener una mejor idea de la relación de los datos con los respectivos factores se puede ver en la figura 7.6 que presión del crono y las tempreaturas de acondicionamiento no reflejan una mayor relación con los factores comunes es decir la rotación no ha sido de mucha ayuda.

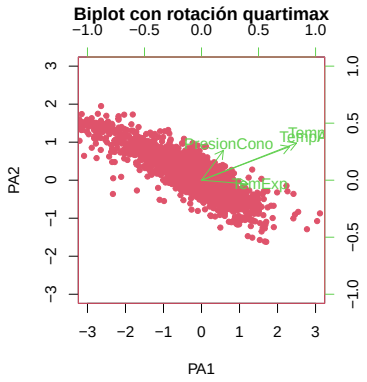

Figura 7.6: Bitplot con rotación quartimax.

7.3.3. Circulo Unitario

El circulo unitario mostrado en 7.7 muestra las relaciones de las variables con su factor común.

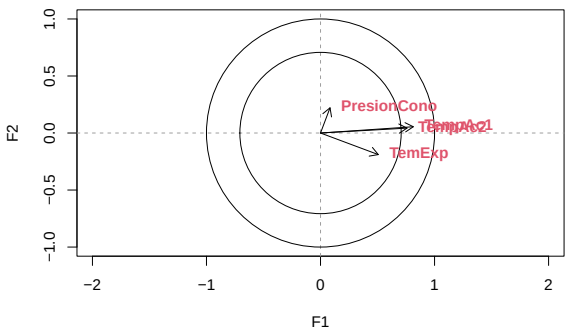

Figura 7.7: Circulo Unitario.

7.3.4. Test de Barlett(KMO)

En el respectivo test de Barllet podemos ver que no hay homogeniedad en la varianza de los grupos como se muestra en la figura 7.8.

```
        Bartlett test of homogeneity of variances

data:  datos2
Bartlett's K-squared = 5939.2, df = 3, p-value < 2.2e-16
```

Figura 7.8: Test de Barlett(KMO)

7.4. Escalado Multidimensional

7.4.1. Diagrama de Sheppard

Podemos notar en el respectivo diagrama de Sheppard que no existe la presencia de un gradiente creciente es decir las distancias y disimilaridades no son concordantes y se altera su relación de orden. En este se obtuvo un valor de Stress de 5.74 corroborando que la configuracion de distancia y disimilaridades no es buena.

```
## initial  value 12.275719
## iter   5 value 10.173015
## iter  10 value 9.339880
## iter  15 value 7.925656
## iter  20 value 5.808498
## final  value 5.740568
## converged
```

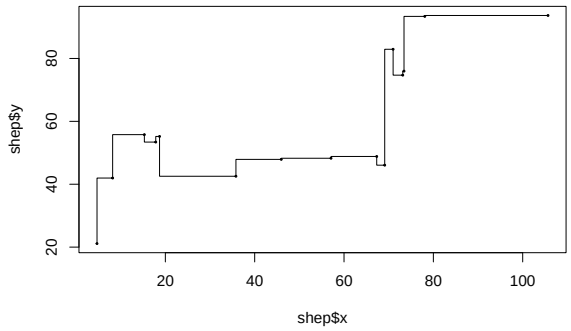

Figura 7.9: Diagrama de Sheppard.

7.4.2. Porcentaje de Varianza Explicada

Se puede ver en la tabla que el 94.3 % de la variabilidad de los datos se representa idoneamente en 4 dimensiones.

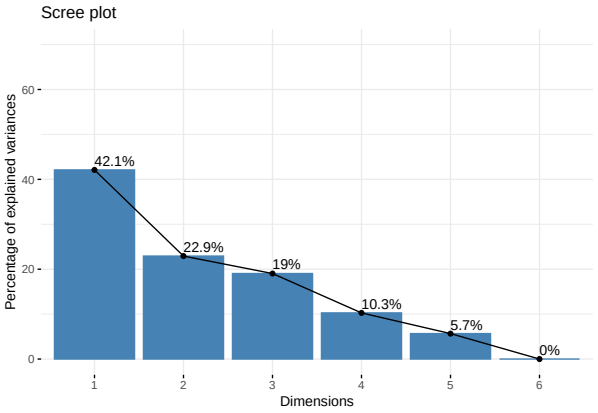

Figura 7.10: Porcentaje de Varianza Explicada.

7.5. Análisis de Correspondencia

7.5.1. Análisis de Correspondencia Múltiple

La figura 7.11 representa la proyección de los puntos fila y columna en el primer plano factorial. Donde las dimensiones son especie (color celeste), operador (color gris), presentación (color verde), proceso (color naranja) y tipo (color morado). Podemos apreciar que todos siguen patrones similares Pero los más próximos que podemos ver es ccondor está considerablemente cerca de ponedora com y pellet.

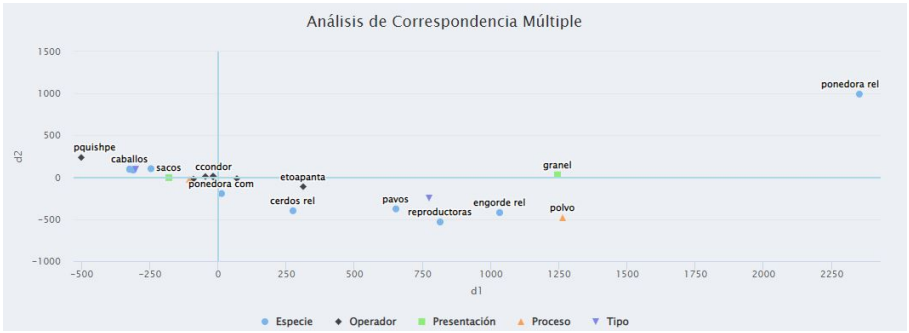

Figura 7.11: Análisis de Correspondencia Múltiple

7.6. Análisis de Conglomerados

7.6.1. Método Vecino más Proximo

A una distancia de 4.5 podemos observar que en la imagen se forman 4 conglomerados, con el método "single", con distancia euclidiana (ver figura 7.12).

```
## [1] 28
```

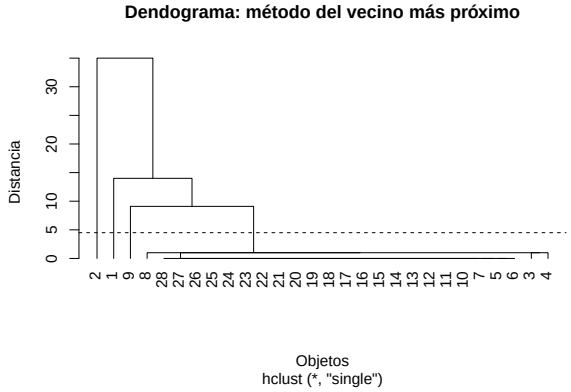

Figura 7.12: Método Vecino más Proximo.

7.6.2. Método Promedio

Con el método “ave” o “average” es difícil poder apreciar los conglomerados porque se forman muchos conglomerados, con distancia euclidiana(ver figura 7.13).

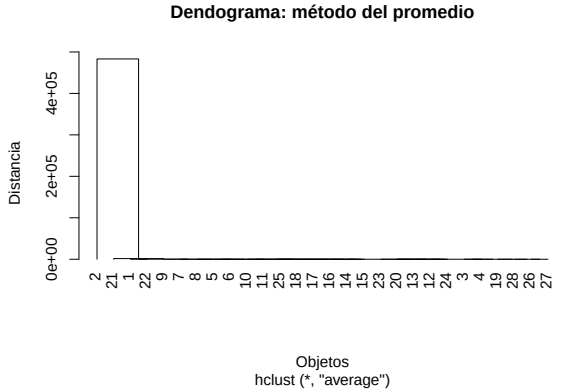

Figura 7.13: Método Promedio.

7.6.3. Método Ward

Con el método “Ward” sucede lo mismo que con el método “average”, que se forman muchos conglomerados y difícil notar los clústeres, con distancia euclidiana(ver figura 7.14).

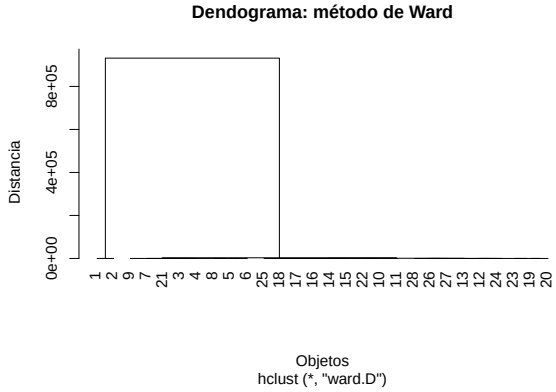

Figura 7.14: Método Ward.

Así mismo podemos probar con diferentes distancias, en este caso se usa la distancia de minkowski para formar los conglomerados, con el método "simple" y estos pueden apreciarse mejor.

7.6.4. Dendograma Cluster

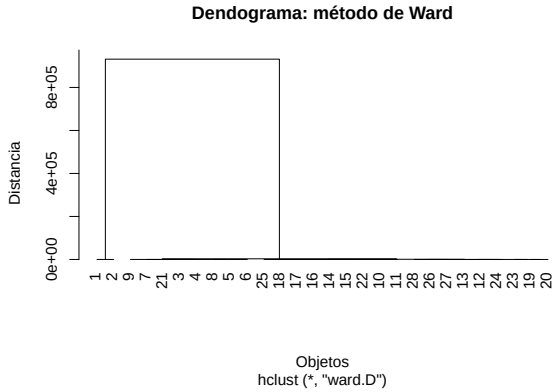

Figura 7.15: Dendograma Cluster.

7.7. Análsis Discrimante

Ayuda a distinguir entre 2 o más grupos de variables y esto resulta de combinaciones lineales de las funciones discriminantes, en el primer grupo (derecha) se aprecia que se agrupan los colores por especies y el de la izquierda son las variables.

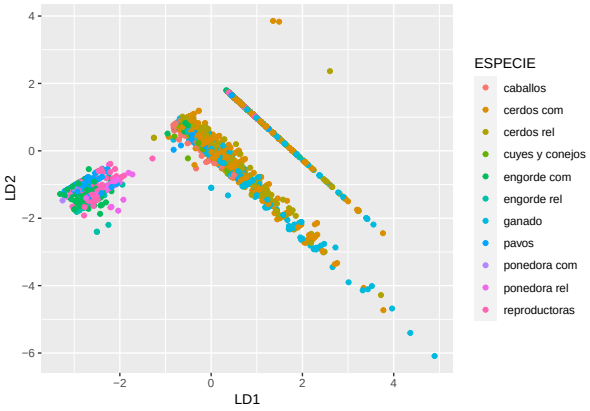

Figura 7.16: Análsis Discrimante.

Capítulo 8

Conclusiones

En el conjunto de datos presentado sobre un proceso de producción de alimento para diferentes tipos de animales de granga para el año 2021, mediante el uso de Estadísticas descriptivas y analisis multivariantes se obtuvieron las siguientes conclusiones:

- Las produccion se maximiza en meses previos a temporadas de consumo masivo de animales de granja en especifico previo a las festividades de diciembre como Navidad y Fin de año.
- Se tienen variables de control en las que se puede evidenciar un trato especifico para cierto producto, por ejemplo la variable Caplinea tiene una alta frecuencia para productos “estrella” provocando que la linea funcione a maxima capacidad.
- Se contrastó que las medias entre temperaturas para cada Maquina difieren.
- Existen variables latentes entre aquellas que se encargan del procesamiento de la materia prima por medio de las variables de control, y aquellas que indican el nivel de produccion como las que miden capacidades en el proceso.
- Sse identificaron conglomerados en los datos especificamente a una distancia de aproximadamente 4.8 se forman 5 grupos.

Se pudieron identificar tanto asociaciones como disociaciones entre las variables es de indicar que en la mayor parte de analisis multivariante 2 dimensiones no era lo adecuado para explicar la variabilidad de los datos sin embargo para una mejor visualizacion del análisis representaciones en 2 dimensiones fue adecuado. Hace falta la opinión de un experto en el proceso de produccion para identificar aquellas variables latentes identificadas.

Apéndice A

Apéndice: Título del Apéndice

A.1. Primera sección

Apéndice B

Apéndice: Título del Apéndice

B.1. Primera sección

Bibliografía

JJ Allaire, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. *rmarkdown: Dynamic Documents for R*, 2022. URL https://CRAN.R-project.org/package=rmarkdown. R package version 2.17.

Pedro L. Luque-Calvo. *Escribir un Trabajo Fin de Estudios con R Markdown*, 2017.

Pedro L. Luque-Calvo. *Cómo crear Tablas de información en R Markdown*, 2019.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL https://www.R-project.org/.

RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA, 2015. URL http://www.rstudio.com/.

Techopedia. "definition - what does business intelligence (bi) mean?". Disponible en https://www.techopedia.com/definition/345/business-intelligence-bi, 2017.

Hadley Wickham, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, and Dewey Dunnington. *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*, 2022a. URL https://CRAN.R-project.org/package=ggplot2. R package version 3.4.0.

Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. *dplyr: A Grammar of Data Manipulation*, 2022b. URL https://CRAN.R-project.org/package=dplyr. R package version 1.0.10.

Yihui Xie. *knitr: A General-Purpose Package for Dynamic Report Generation in R*, 2022. URL https://yihui.org/knitr/. R package version 1.40.