



# UNIVERSIDAD DE GRANADA

Grado en Estadística  
(Curso 2018-2019)

## Trabajo Fin de Grado

**Análisis Estadístico de Datos Textuales. Una aproximación exploratoria.**

*J. David Fernández Romero  
(Tutor: José Fernando Vera Vera)*

## Sumario

- Breve historia del análisis estadístico de *datos textuales*.
  - Origen.
  - Evolución.

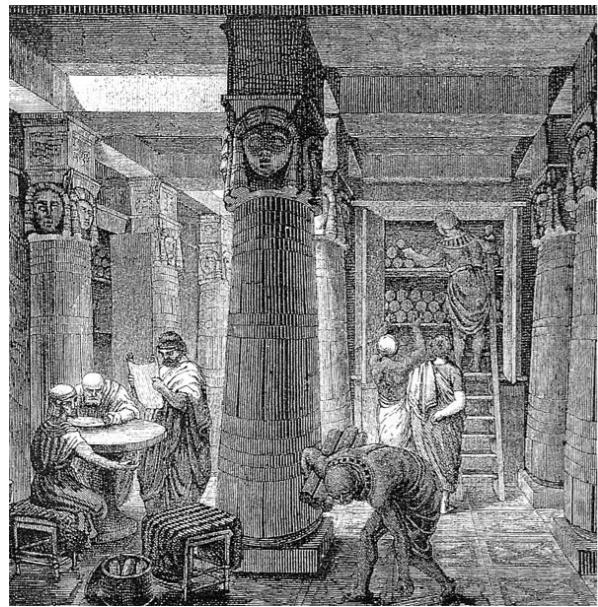
# Sumario

- Breve historia del análisis estadístico de *datos textuales*.
  - Origen.
  - Evolución.
- “Wine Reviews”
  - > 100.000 reseñas de vinos.
  - Variables.
    - Cualitativas
    - Cuantitativas
    - “Descripción”

# Sumario

- Breve historia del análisis estadístico de *datos textuales*.
  - Origen.
  - Evolución.
- “Wine Reviews”
  - > 100.000 reseñas de vinos.
  - Variables.
    - Cualitativas
    - Cuantitativas
    - “Descripción”
  - Variable textual.
    - Preprocesado. Codificación.
    - Expresiones frecuentes. Concordancias. Expresiones características.
    - Análisis de correspondencias.

# Evolución del análisis estadístico de datos textuales

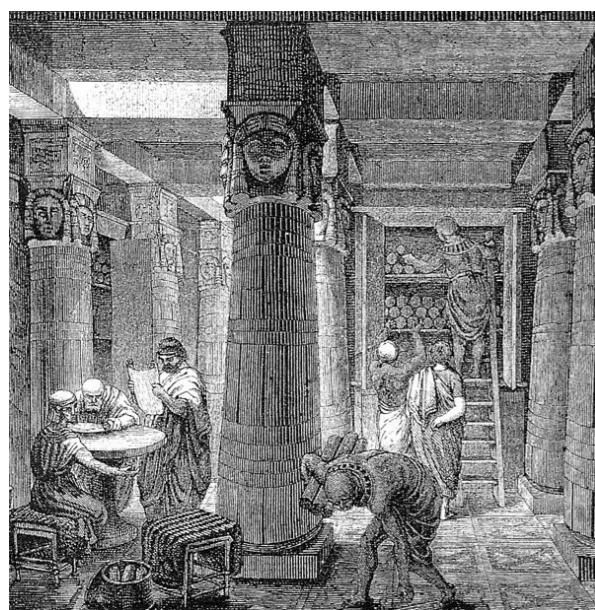


- **Inventario de palabras de la Biblia**
- **Hapax de Homero** (*hapax*: término que cuenta con una única amación en un texto)

# Evolución del análisis estadístico de datos textuales

1930 - ...: Leyes empíricas sobre la distribución de las palabras

- (Zipf, Yule, Guiraud, ...)

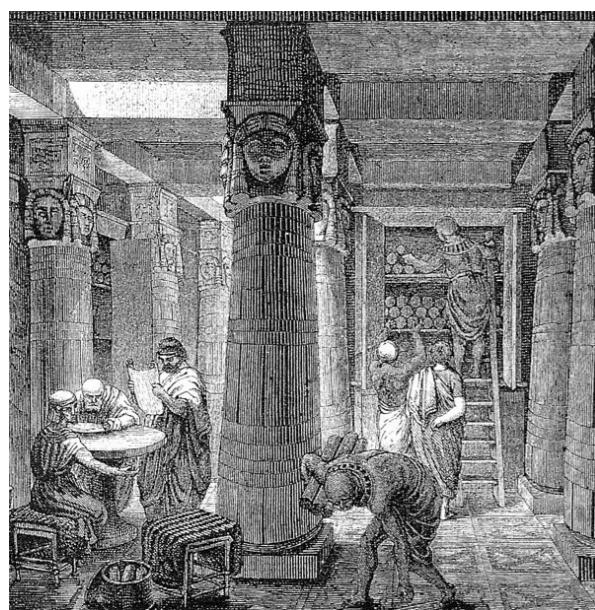


- **Inventario de palabras de la Biblia**
- **Hapax de Homero** (*hapax*: término que cuenta con una única amación en un texto)

# Evolución del análisis estadístico de datos textuales

**1930 - ...: Leyes empíricas sobre la distribución de las palabras**  
• (Zipf, Yule, Guiraud, ...)

**1964: Benzécri - “Lingüística matemática”**  
• Grandes tablas de datos  
• Aplicación en muchos campos



- **Inventario de palabras de la Biblia**
- **Hapax de Homero** (*hapax*: término que cuenta con una única amación en un texto)

# Evolución del análisis estadístico de datos textuales

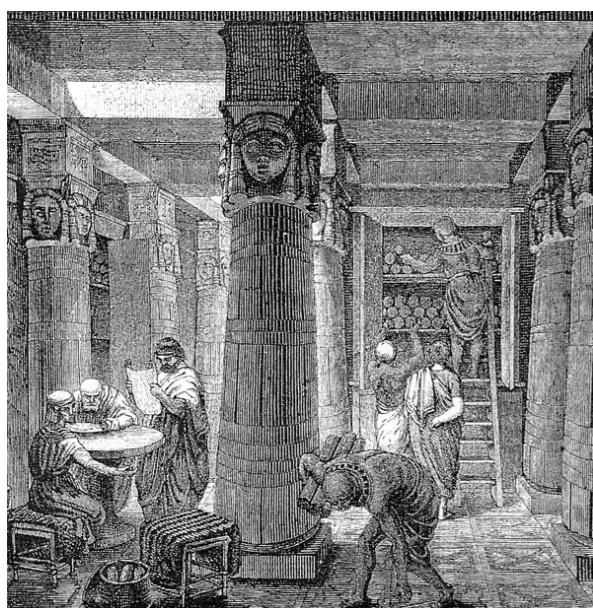
**1930 - ...: Leyes empíricas sobre la distribución de las palabras**  
• (Zipf, Yule, Guiraud, ...)

**1964: Benzécri - “Lingüística matemática”**  
• Grandes tablas de datos  
• Aplicación en muchos campos

**1980 ...**

**Fuerte desarrollo de la informática**  
• Hardware: Aumento de la capacidad de almacenamiento y procesamiento  
• Software: Desarrollo de programas más potentes

**Análisis multivariante**



- **Inventario de palabras de la Biblia**
- **Hapax de Homero** (*hapax*: término que cuenta con una única amación en un texto)

# Evolución del análisis estadístico de datos textuales

**1930 - ...: Leyes empíricas sobre la distribución de las palabras**  
• (Zipf, Yule, Guiraud, ...)

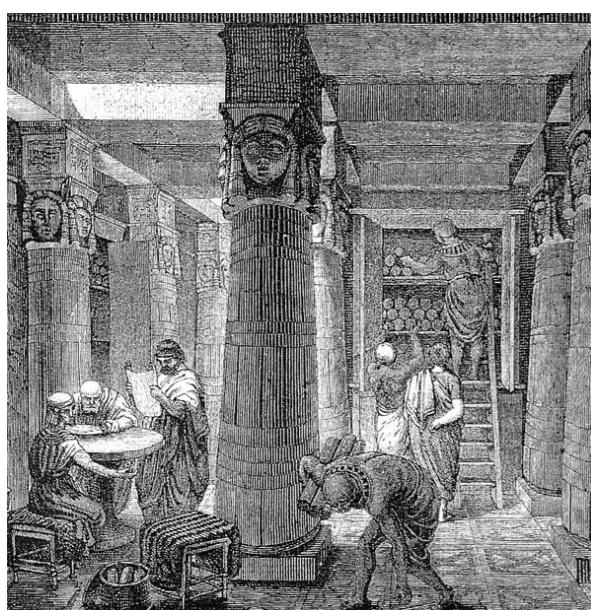
**1984: L. Lebart / A. Morineau**  
• Módulo de tratamiento de textos en SPAD

**1964: Benzécri - “Lingüística matemática”**  
• Grandes tablas de datos  
• Aplicación en muchos campos

**1980 ...**

**Fuerte desarrollo de la informática**  
• Hardware: Aumento de la capacidad de almacenamiento y procesamiento  
• Software: Desarrollo de programas mas potentes

**Análisis multivariante**



- **Inventario de palabras de la Biblia**
- **Hapax de Homero** (*hapax*: término que cuenta con una única amación en un texto)

# Evolución del análisis estadístico de datos textuales

**1930 - ...: Leyes empíricas sobre la distribución de las palabras**  
• (Zipf, Yule, Guiraud, ...)

**1984: L. Lebart / A. Morineau**  
• Módulo de tratamiento de textos en SPAD

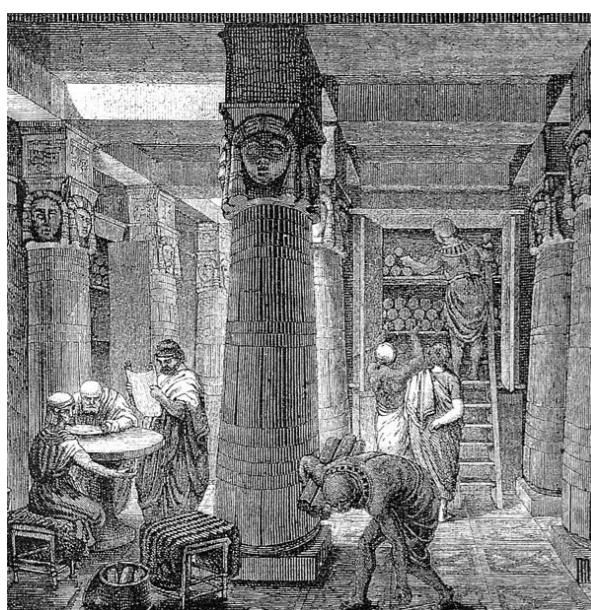
**1964: Benzécri - “Lingüística matemática”**  
• Grandes tablas de datos  
• Aplicación en muchos campos

**1988: M. Bécue Bertaut - SPAD.T**

**1980 ...**

**Fuerte desarrollo de la informática**  
• Hardware: Aumento de la capacidad de almacenamiento y procesamiento  
• Software: Desarrollo de programas mas potentes

**Análisis multivariante**



- **Inventario de palabras de la Biblia**
- **Hapax de Homero** (*hapax*: término que cuenta con una única amación en un texto)

# Evolución del análisis estadístico de datos textuales

**1930 - ...: Leyes empíricas sobre la distribución de las palabras**  
• (Zipf, Yule, Guiraud, ...)

**1984: L. Lebart / A. Morineau**  
• Módulo de tratamiento de textos en SPAD

**1964: Benzécri - “Lingüística matemática”**  
• Grandes tablas de datos  
• Aplicación en muchos campos

**1988: M. Bécue Bertaut - SPAD.T**

**1980 ...**

**Fuerte desarrollo de la informática**  
• Hardware: Aumento de la capacidad de almacenamiento y procesamiento  
• Software: Desarrollo de programas mas potentes

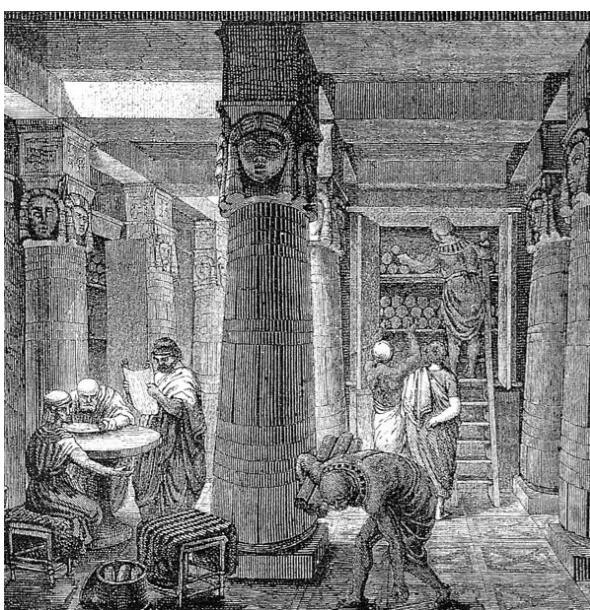
**Análisis multivariante**

**2010 ...**

**Minería de textos**

• Redes Sociales  
• Smartphones  
• Big Data

- **Inventario de palabras de la Biblia**
- **Hapax de Homero** (*hapax*: término que cuenta con una única amación en un texto)



## Wine reviews

- **Variables**

- **Cualitativas:**

- **País / Provincia / Región 1 / Región 2**
    - **Variedad**
    - **Bodega / Nombre comercial**

## Wine reviews

- **Variables**

- **Cualitativas:**

- País / Provincia / Región 1 / Región 2
    - Variedad
    - Bodega / Nombre comercial

- **Cuantitativas:**

- Puntuación
    - Precio

# Wine reviews

- **Variables**

- **Cualitativas:**

- País / Provincia / Región 1 / Región 2
    - Variedad
    - Bodega / Nombre comercial

- **Cuantitativas:**

- Puntuación
    - Precio

- **Textual:**

- Descripción.

# Wine reviews

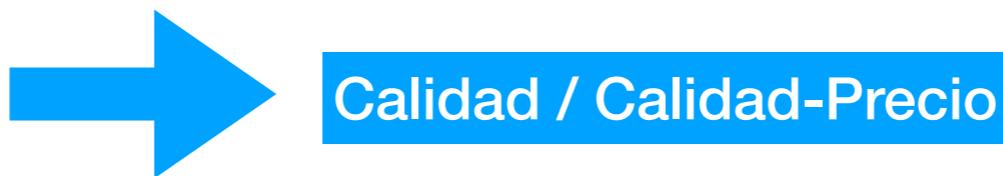
- **Variables**

- **Cualitativas:**

- País / Provincia / Región 1 / Región 2
    - Variedad
    - Bodega / Nombre comercial

- **Cuantitativas:**

- Puntuación
    - Precio



- **Textual:**

- Descripción.

# Wine reviews

- **Variables**

- **Cualitativas:**

- País / Provincia / Región 1 / Región 2
    - Variedad
    - Bodega / Nombre comercial

- **Cuantitativas:**

- Puntuación
    - Precio

Calidad / Calidad-Precio

- **Textual:**

- Descripción.

NO directamente tratable  
mediante técnicas estadísticas



PREPROCESADO



CODIFICACIÓN



TABLA MULTIVARIANTE

# Variable textual - Preprocesado

## DESCRIPCION

- 1 To get the best results, you'd better chill this one down. That way you'll get more of the refreshing apple and pear element of the wine and less of its fat, chunky body and rough, unpolished edges.
- 2 This yellow-colored blend is weird and wooly, with aromas of blowsy apricot and herbs. A pulpy, melony palate is creamy and odd in feel, with nectarine and chunky peach flavors that turn rubbery on the finish.
- 3 Yeasty, melony and smelling like yellow mustard, this Moscato is generally unappealing. Spritzy (but barely) on the palate, it has flavors of canned pineapple and mango, with a cloying finish.
- 4 While this smells floral and candied, the palate rages, courtesy of strapping acidity that creates tart plum and blueberry flavors.
- 5 Tasted in early February, this Chardonnay had some secondary fermentation in the bottle, giving it a bubbly feeling. This major flaw may disappear in time.
- 6 These vineyards seem to have been selected for unripe, vegetal grapes. Just manages to scrape by with a veneer of cherries.

# Variable textual - Preprocesado

## DESCRIPCION

- 1 To get the best results, you'd better chill this one down. That way you'll get more of the refreshing apple and pear element of the wine and less of its fat, chunky body and rough, unpolished edges.
- 2 This yellow-colored blend is weird and wooly, with aromas of blowsy apricot and herbs. A pulpy, melony palate is creamy and odd in feel, with nectarine and chunky peach flavors that turn rubbery on the finish.
- 3 Yeasty, melony and smelling like yellow mustard, this Moscato is generally unappealing. Spritzy (but barely) on the palate, it has flavors of canned pineapple and mango, with a cloying finish.
- 4 While this smells floral and candied, the palate rages, courtesy of strapping acidity that creates tart plum and blueberry flavors.
- 5 Tasted in early February, this Chardonnay had some secondary fermentation in the bottle, giving it a bubbly feeling. This major flaw may disappear in time.
- 6 These vineyards seem to have been selected for unripe, vegetal grapes. Just manages to scrape by with a veneer of cherries.

- **Corrección ortográfica + limpieza de notaciones normativas + signos de puntuación**
- **Segmentación en unidades léxicas:** **Palabra** (*Secuencia de caracteres entre espacios / puntuación*)      **Lema** (*Entrada del diccionario asociada a una palabra*)
  - **Unidades léxicas llenas:** aportan significado (sustantivos, verbos, ...)
  - **Unidades léxicas gramaticales:** no aportan significado (artículos, preposiciones, conjunciones, ...)

# Variable textual - Preprocesado

## DESCRIPCION

- 1 To get the best results, you'd better chill this one down. That way you'll get more of the refreshing apple and pear element of the wine and less of its fat, chunky body and rough, unpolished edges.
- 2 This yellow-colored blend is weird and wooly, with aromas of blowsy apricot and herbs. A pulpy, melony palate is creamy and odd in feel, with nectarine and chunky peach flavors that turn rubbery on the finish.
- 3 Yeasty, melony and smelling like yellow mustard, this Moscato is generally unappealing. Spritzy (but barely) on the palate, it has flavors of canned pineapple and mango, with a cloying finish.
- 4 While this smells floral and candied, the palate rages, courtesy of strapping acidity that creates tart plum and blueberry flavors.
- 5 Tasted in early February, this Chardonnay had some secondary fermentation in the bottle, giving it a bubbly feeling. This major flaw may disappear in time.
- 6 These vineyards seem to have been selected for unripe, vegetal grapes. Just manages to scrape by with a veneer of cherries.

- **Corrección ortográfica + limpieza de notaciones normativas + signos de puntuación**
- **Segmentación en unidades léxicas:** **Palabra** (*Secuencia de caracteres entre espacios / puntuación*)      **Lema** (*Entrada del diccionario asociada a una palabra*)
  - **Unidades léxicas llenas:** aportan significado (sustantivos, verbos, ...)
  - **Unidades léxicas gramaticales:** no aportan significado (artículos, preposiciones, conjunciones, ...)

- **Segmentos repetidos**
- **“Stematización”**
- **Umbral de frecuencia**

# Variable textual - Preprocesado

## DESCRIPCION

- 1 To get the best results, you'd better chill this one down. That way you'll get more of the refreshing apple and pear element of the wine and less of its fat, chunky body and rough, unpolished edges.
- 2 This yellow-colored blend is weird and wooly, with aromas of blowsy apricot and herbs. A pulpy, melony palate is creamy and odd in feel, with nectarine and chunky peach flavors that turn rubbery on the finish.
- 3 Yeasty, melony and smelling like yellow mustard, this Moscato is generally unappealing. Spritzy (but barely) on the palate, it has flavors of canned pineapple and mango, with a cloying finish.
- 4 While this smells floral and candied, the palate rages, courtesy of strapping acidity that creates tart plum and blueberry flavors.
- 5 Tasted in early February, this Chardonnay had some secondary fermentation in the bottle, giving it a bubbly feeling. This major flaw may disappear in time.
- 6 These vineyards seem to have been selected for unripe, vegetal grapes. Just manages to scrape by with a veneer of cherries.

- **Corrección ortográfica + limpieza de notaciones normativas + signos de puntuación**
- **Segmentación en unidades léxicas:** **Palabra** (*Secuencia de caracteres entre espacios / puntuación*)      **Lema** (*Entrada del diccionario asociada a una palabra*)
  - **Unidades léxicas llenas:** aportan significado (sustantivos, verbos, ...)
  - **Unidades léxicas gramaticales:** no aportan significado (artículos, preposiciones, conjunciones, ...)

- **Segmentos repetidos**
- **“Stematización”**
- **Umbral de frecuencia**

## DESCRIPCION

- 1 get best results better chill one way get refreshing apple pear element wine less fat chunky body rough unpolished edges
- 2 yellowcolored blend weird wooly aromas blowsy apricot herbs pulpy melony palate creamy odd feel nectarine chunky peach flavors turn rubbery finish
- 3 yeasty melony smelling like yellow mustard moscato generally unappealing spritzy barely palate flavors canned pineapple mango cloying finish
- 4 smells floral candied palate rages courtesy strapping acidity creates tart plum blueberry flavors
- 5 tasted early february chardonnay secondary fermentation bottle giving bubbly feeling major flaw may disappear time
- 6 vineyards seem selected unripe vegetal grapes just manages scrape veneer cherries

## Variable textual - Codificación

- **Tabla léxica (Documentos x Palabras):** Tabla de frecuencias que contiene el número de veces que en cada documento aparece cada una de las palabras

Docs	Terms									
	acidity	aromas	cherrv	drink	finish	flavors	fruit	palate	tannins	wine
109636	0	0	0	0	2	2	0	0	2	0
118787	1	1	0	1	1	1	2	1	0	0
119835	0	0	1	0	1	1	2	1	0	0
120527	1	0	0	0	0	0	0	1	0	2
120648	0	0	0	1	1	0	1	1	0	1
120685	0	0	0	0	0	0	1	0	0	0
120883	0	0	0	0	0	1	0	0	0	1
120903	0	0	0	0	0	1	0	0	1	3
47255	1	1	0	0	1	0	0	1	0	2
97891	0	0	2	0	2	0	2	2	2	1

## Variable textual - Codificación

- **Tabla léxica (Documentos x Palabras):** Tabla de frecuencias que contiene el número de veces que en cada documento aparece cada una de las palabras

Docs	Terms									
	acidity	aromas	cherrv	drink	finish	flavors	fruit	palate	tannins	wine
109636	0	0	0	0	2	2	0	0	2	0
118787	1	1	0	1	1	1	2	1	0	0
119835	0	0	1	0	1	1	2	1	0	0
120527	1	0	0	0	0	0	0	1	0	2
120648	0	0	0	1	1	0	1	1	0	1
120685	0	0	0	0	0	0	1	0	0	0
120883	0	0	0	0	0	1	0	0	0	1
120903	0	0	0	0	0	1	0	0	1	3
47255	1	1	0	0	1	0	0	1	0	2
97891	0	0	2	0	2	0	2	2	2	1

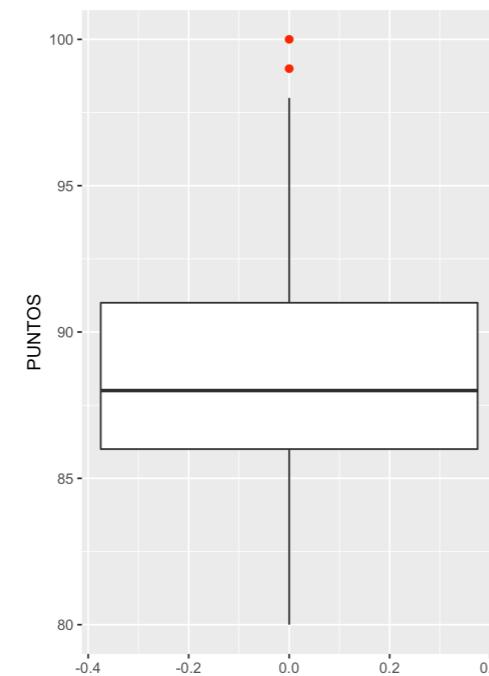
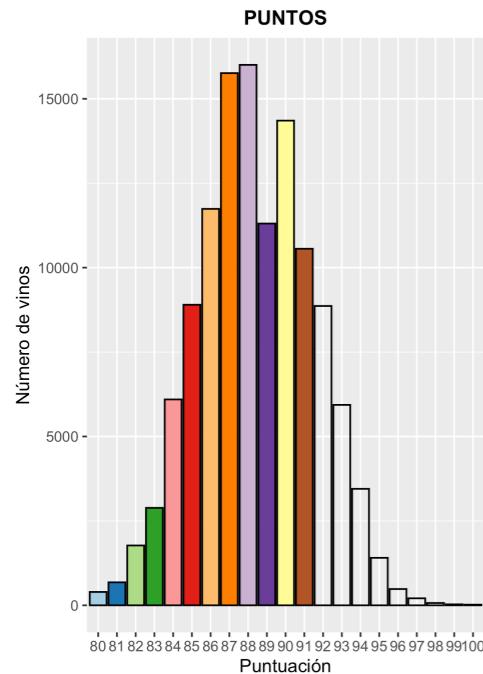
- **Tabla léxica agregada (Palabras x Categorías):** Se agregan los documentos por una tercera variable categórica y se construye la tabla de frecuencias que contiene el número de veces que en cada categoría aparece cada una de las palabras

Docs	Terms									
	acidity	aromas	cherry	drink	finish	flavors	fruit	palate	ripe	tannins
PTOS_MED - BARATO	5514	5910	2993	4550	5921	9782	6295	5793	4505	3515
PTOS_MED - CARO	3432	4874	4778	3617	3689	6611	5696	4792	3267	5206
PTOS_MED - PRECIO_MED	9463	11460	7513	7256	9637	16771	12444	10873	7296	8234
Q1_PTOS - BARATO	2334	3522	1400	1797	3093	5485	2437	2444	1070	1075
Q1_PTOS - CARO	1269	1615	1758	1129	1758	4101	2421	1685	1009	1700
Q1_PTOS - PRECIO_MED	4021	4985	2486	2672	4566	8486	4602	4136	1979	2203
Q4_PTOS - BARATO	1534	1160	921	1573	1420	2418	2038	1977	1415	1288
Q4_PTOS - CARO	1210	1311	1233	1927	1118	1958	1799	1304	1327	1936
Q4_PTOS - PRECIO_MED	2736	2629	2884	3121	2452	4522	3964	3367	2440	2954

# Puntuación / Precio

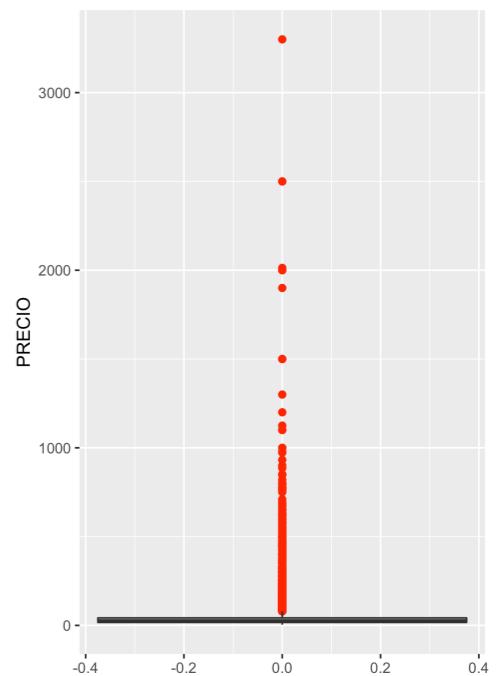
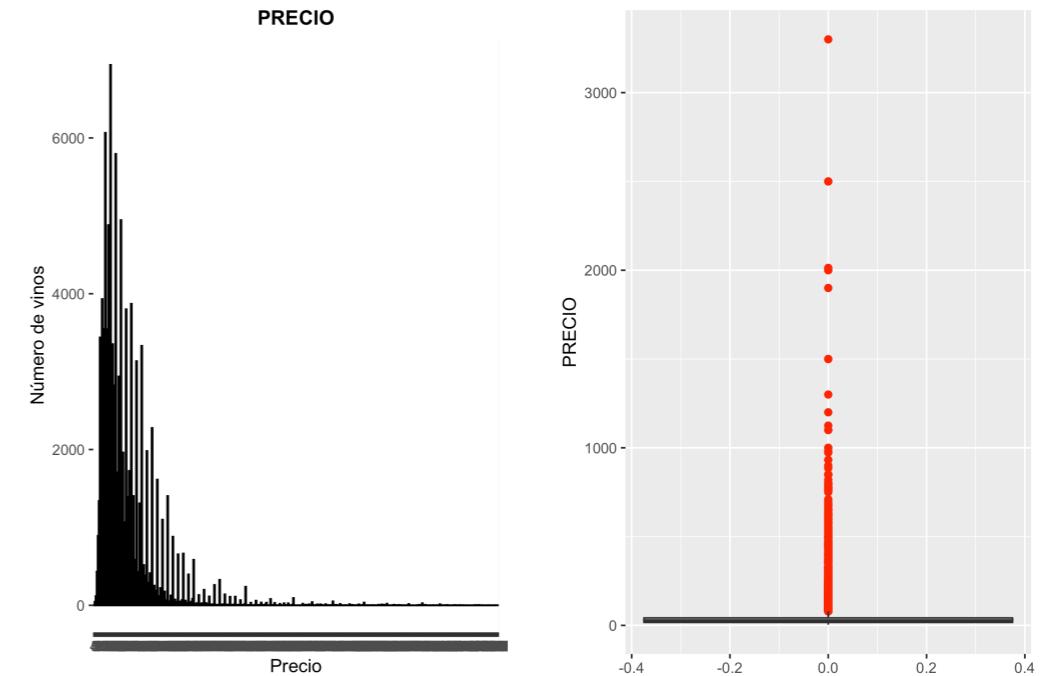
## "PUNTOS"

Min. 1st Qu. Median Mean 3rd Qu. Max.  
80.00 86.00 88.00 88.42 91.00 100.00



## "PRECIO"

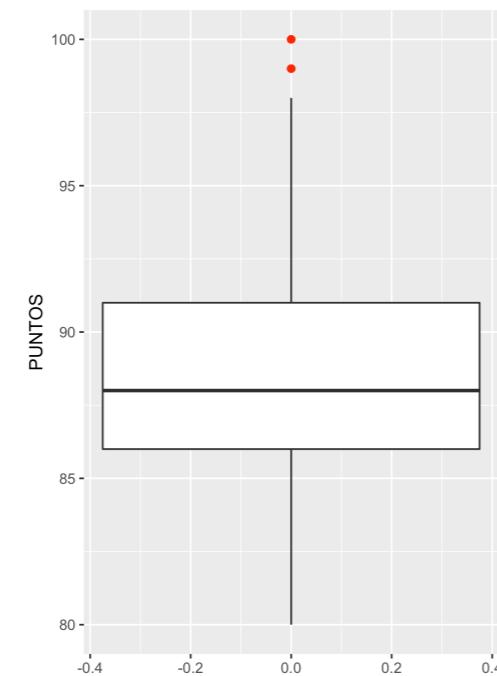
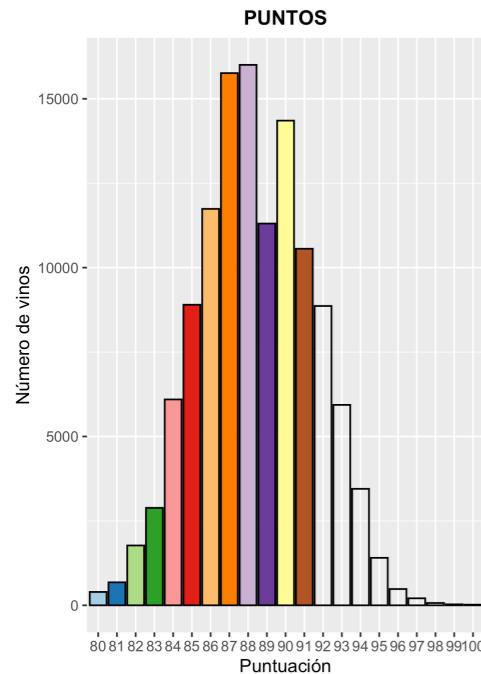
Min. 1st Qu. Median Mean 3rd Qu. Max.  
4.00 17.00 25.00 35.37 42.00 3300.00



# Puntuación / Precio

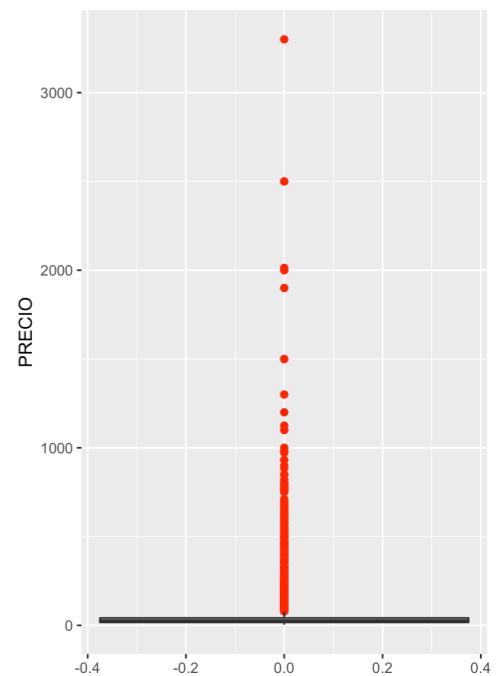
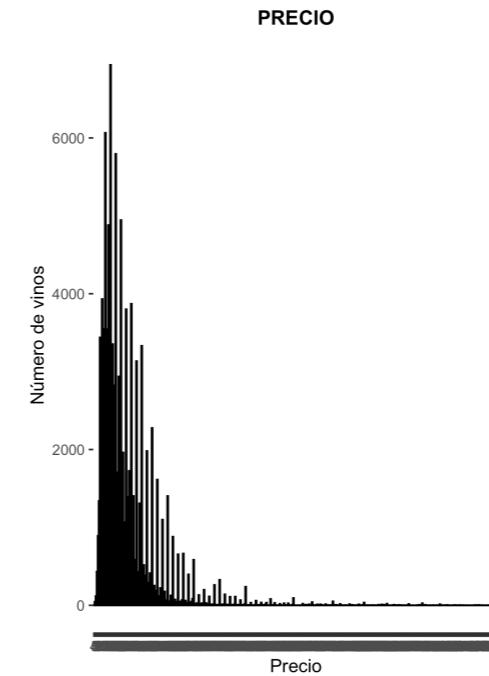
## "PUNTOS"

Min. 1st Qu. Median Mean 3rd Qu. Max.  
 80.00 86.00 88.00 88.42 91.00 100.00



## "PRECIO"

Min. 1st Qu. Median Mean 3rd Qu. Max.  
 4.00 17.00 25.00 35.37 42.00 3300.00



## RANGO\_PUNTOS

**Q1\_PTOS:** 80-86 puntos  
**PTOS\_MED:** 87-91 puntos  
**Q4\_PTOS:** 92-100 puntos

## RANGO\_PRECIO

(vinos con igual puntuación)

**BARATO:** Q1 precios

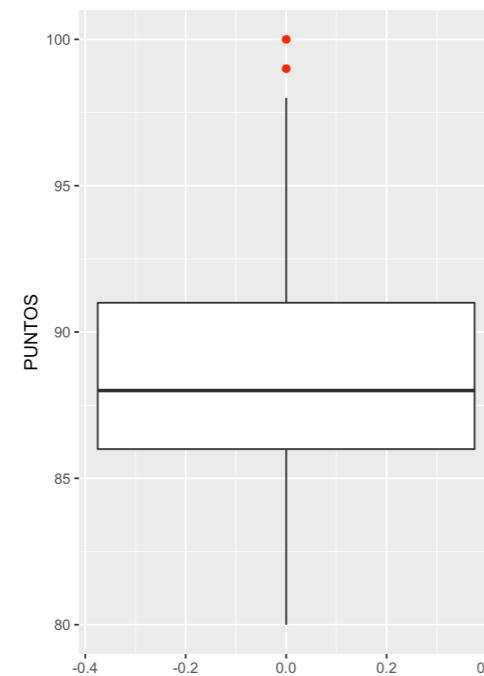
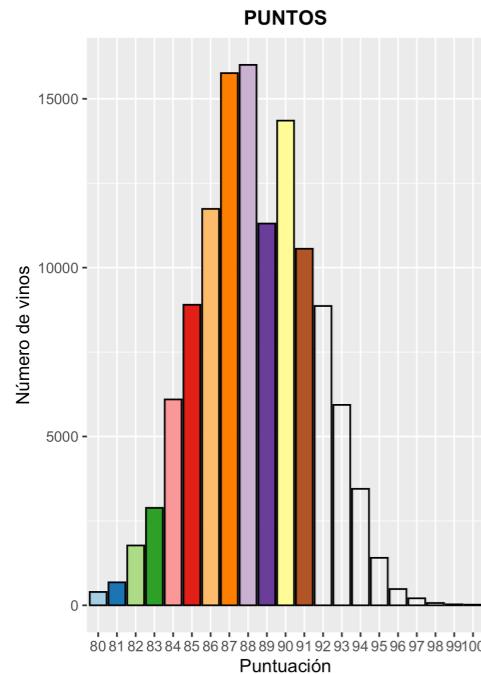
**PRECIO\_MED:** Q2+Q3 precios

**CARO:** Q4 precios

# Puntuación / Precio

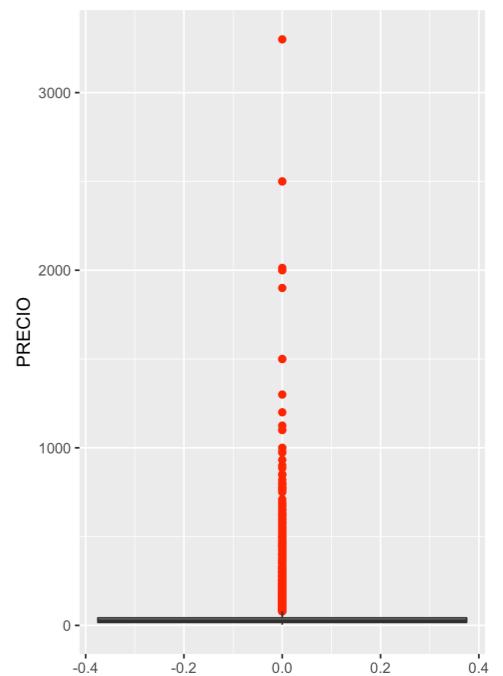
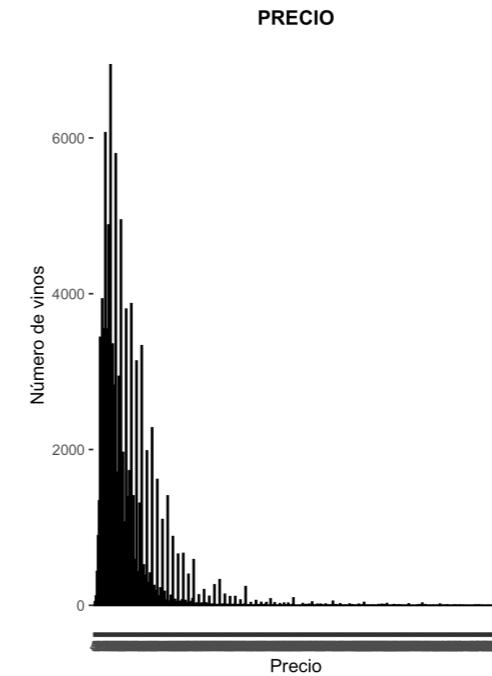
## "PUNTOS"

Min. 1st Qu. Median Mean 3rd Qu. Max.  
 80.00 86.00 88.00 88.42 91.00 100.00



## "PRECIO"

Min. 1st Qu. Median Mean 3rd Qu. Max.  
 4.00 17.00 25.00 35.37 42.00 3300.00



## RANGO\_PUNTOS

Q1\_PTOS: 80-86 puntos  
 PTOS\_MED: 87-91 puntos  
 Q4\_PTOS: 92-100 puntos

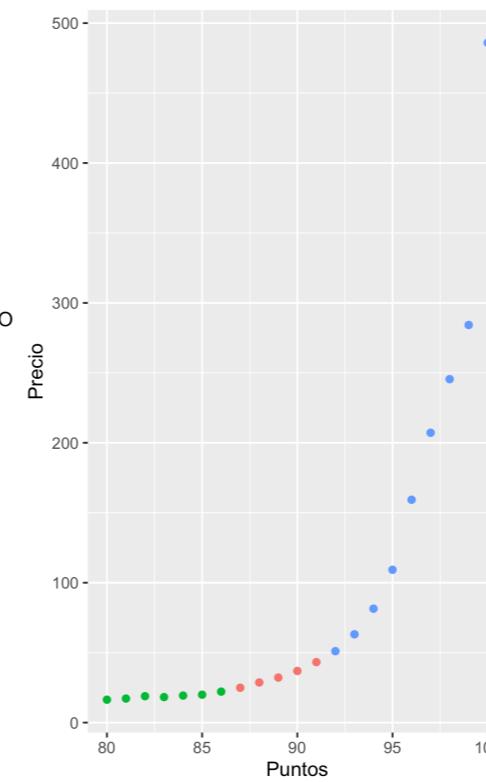
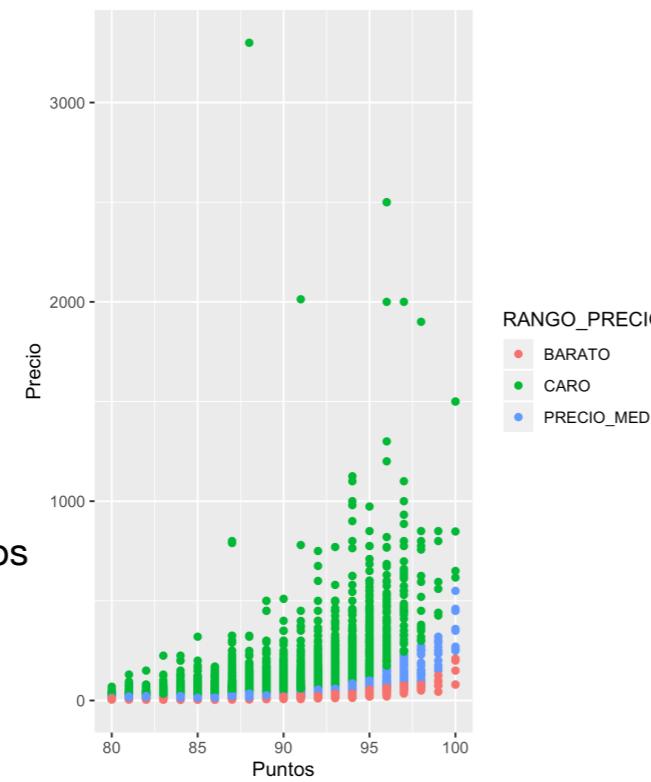
## RANGO\_PRECIO

(vinos con igual puntuación)

**BARATO:** Q1 precios

**PRECIO\_MED:** Q2+Q3 precios

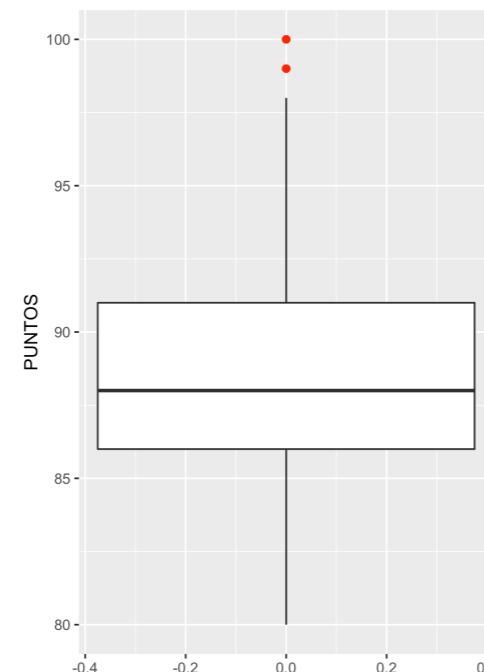
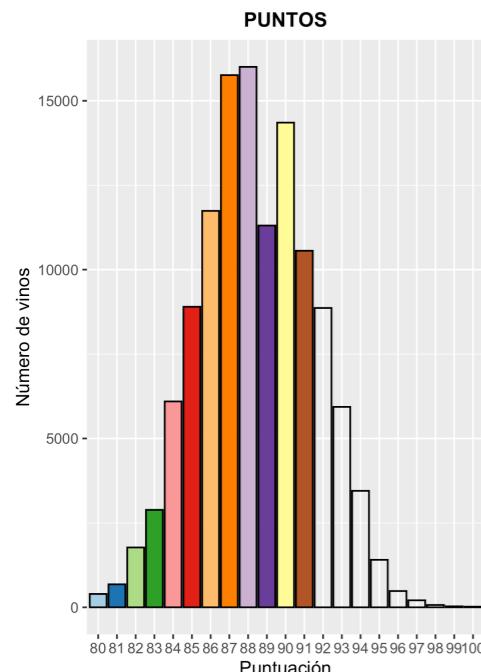
**CARO:** Q4 precios



# Puntuación / Precio

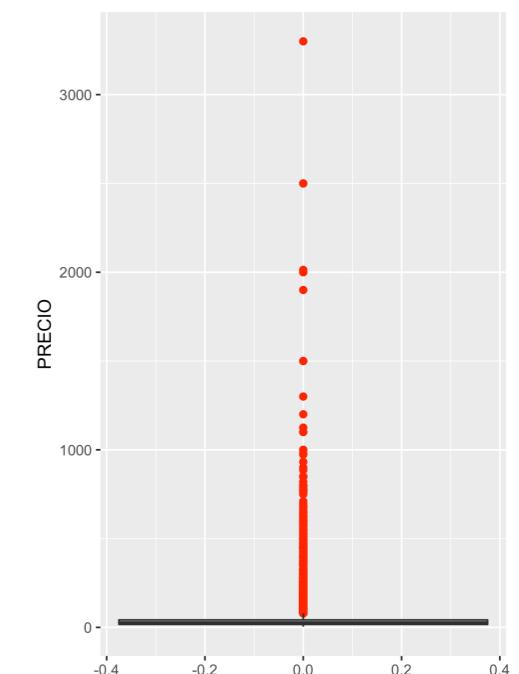
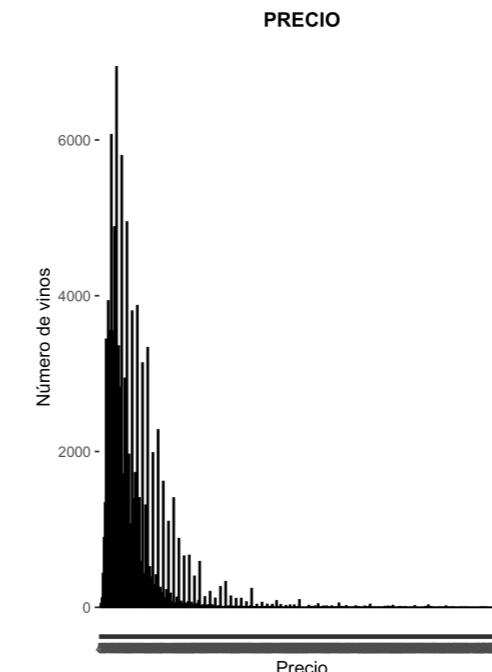
## "PUNTOS"

Min. 1st Qu. Median Mean 3rd Qu. Max.  
 80.00 86.00 88.00 88.42 91.00 100.00



## "PRECIO"

Min. 1st Qu. Median Mean 3rd Qu. Max.  
 4.00 17.00 25.00 35.37 42.00 3300.00



## RANGO\_PUNTOS

Q1\_PTOS: 80-86 puntos  
 PTOS\_MED: 87-91 puntos  
 Q4\_PTOS: 92-100 puntos

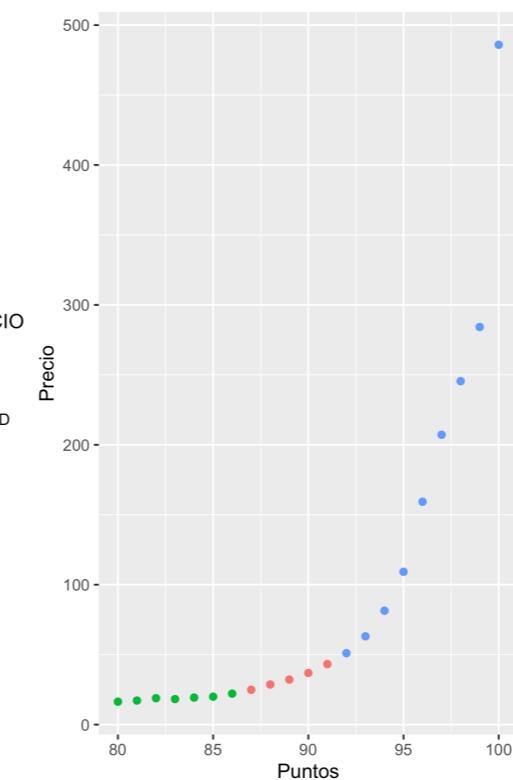
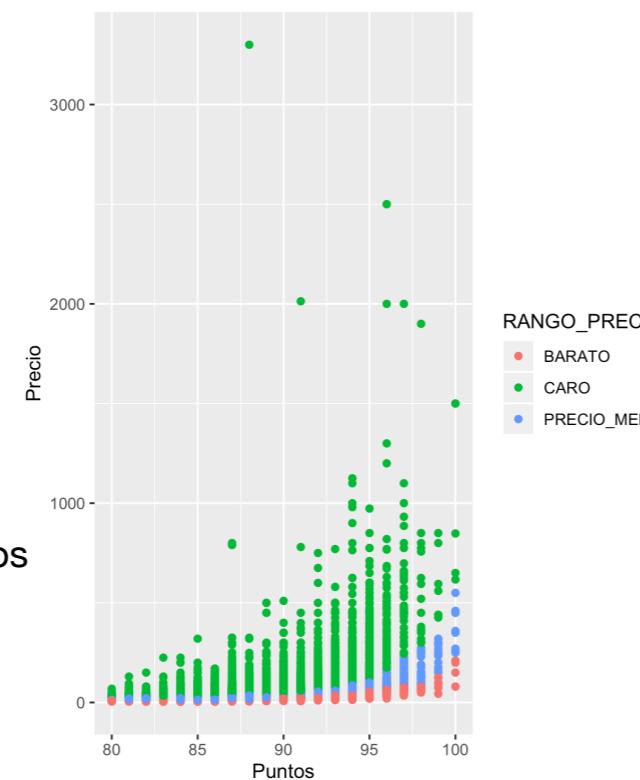
## RANGO\_PRECIO

(vinos con igual puntuación)

**BARATO:** Q1 precios

**PRECIO\_MED:** Q2+Q3 precios

**CARO:** Q4 precios



## "PUNTOS - PRECIO"

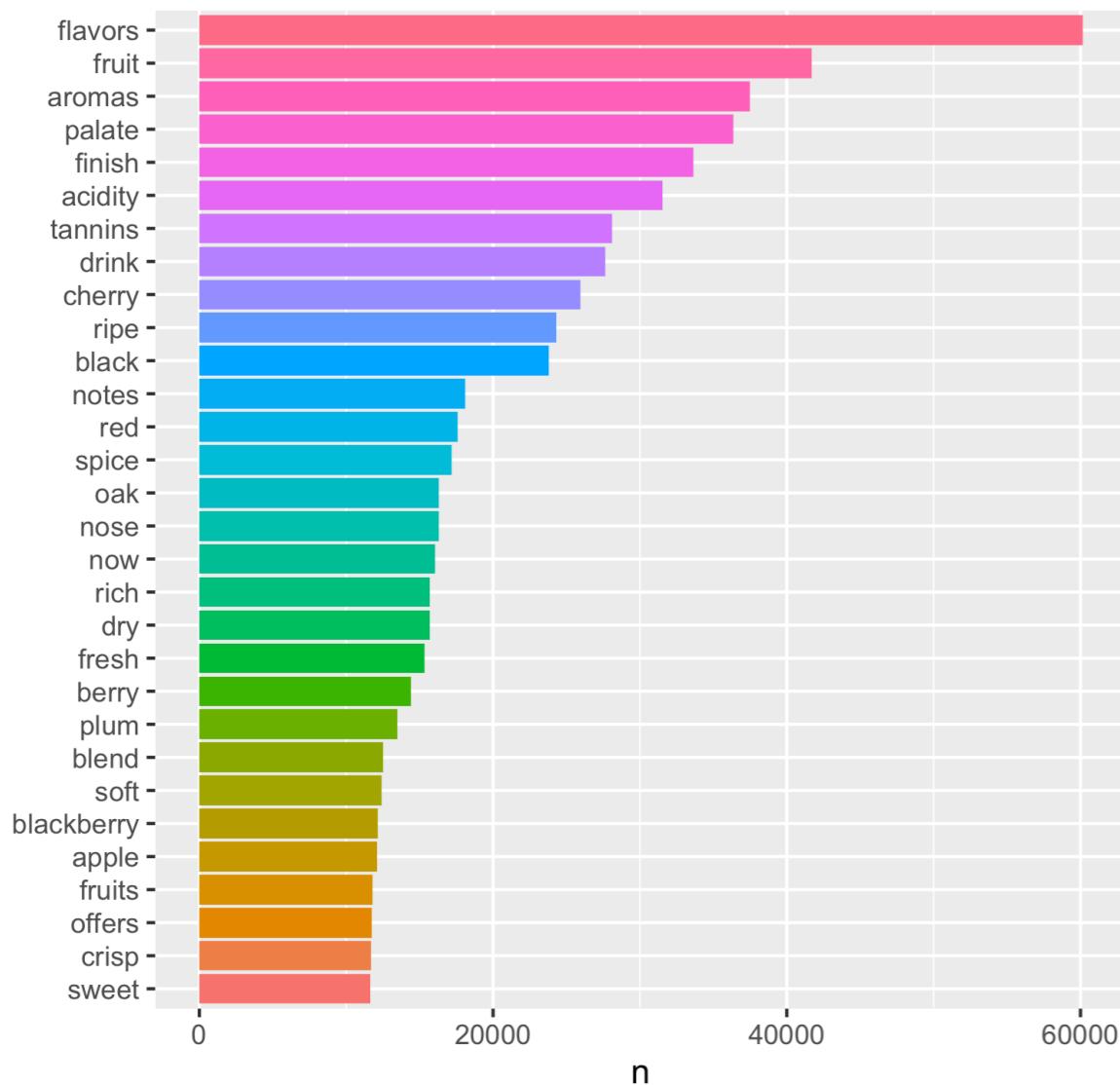
"COVARIANZA"  
 51.99851  
 "CORRELACIÓN"  
 0.4161973

## "PUNTOS - PRECIO MEDIO"

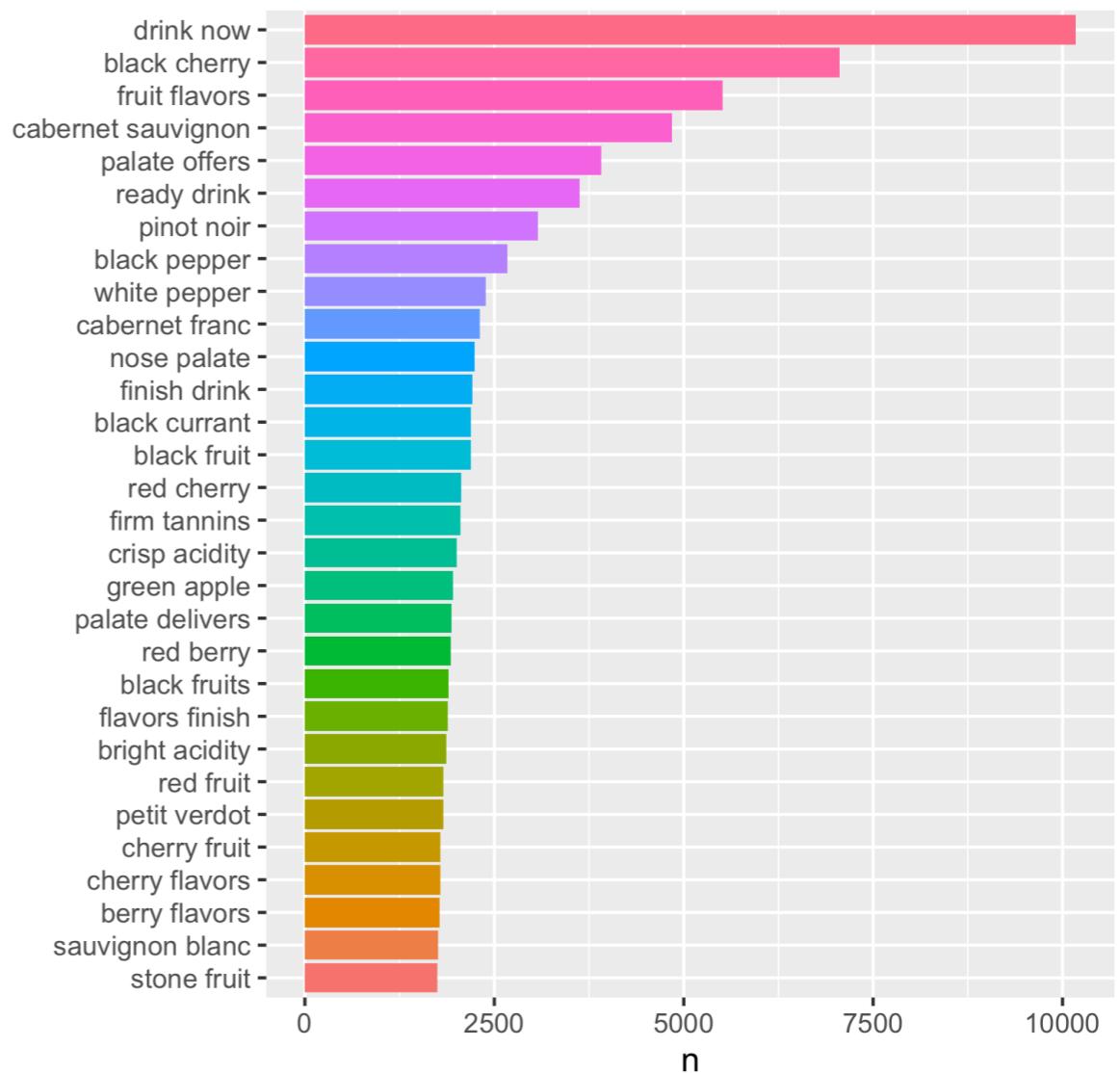
"COVARIANZA"  
 596.4234  
 "CORRELACIÓN"  
 0.7965461

# Expresiones frecuentes

Términos más utilizados

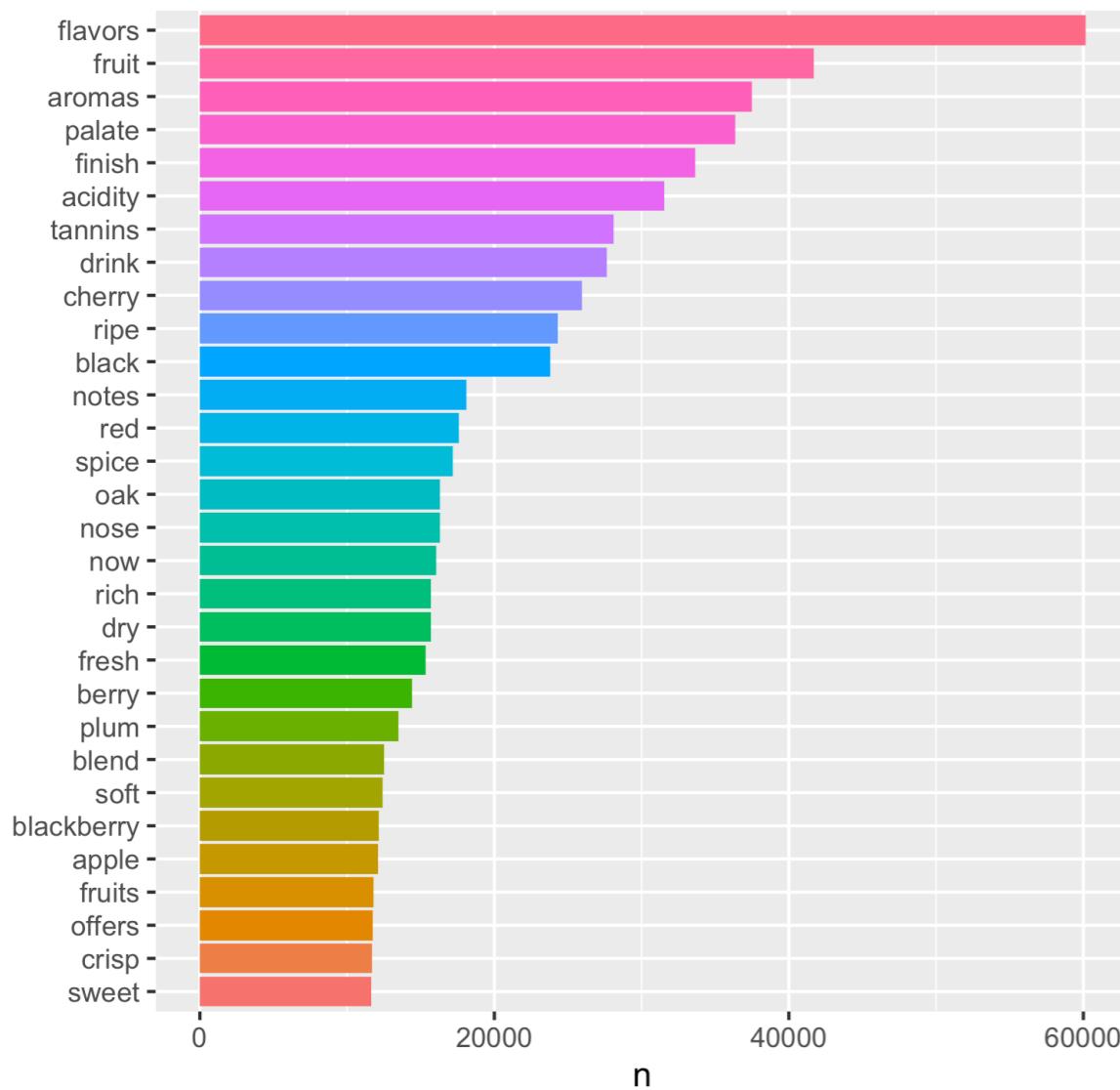


Bigramas

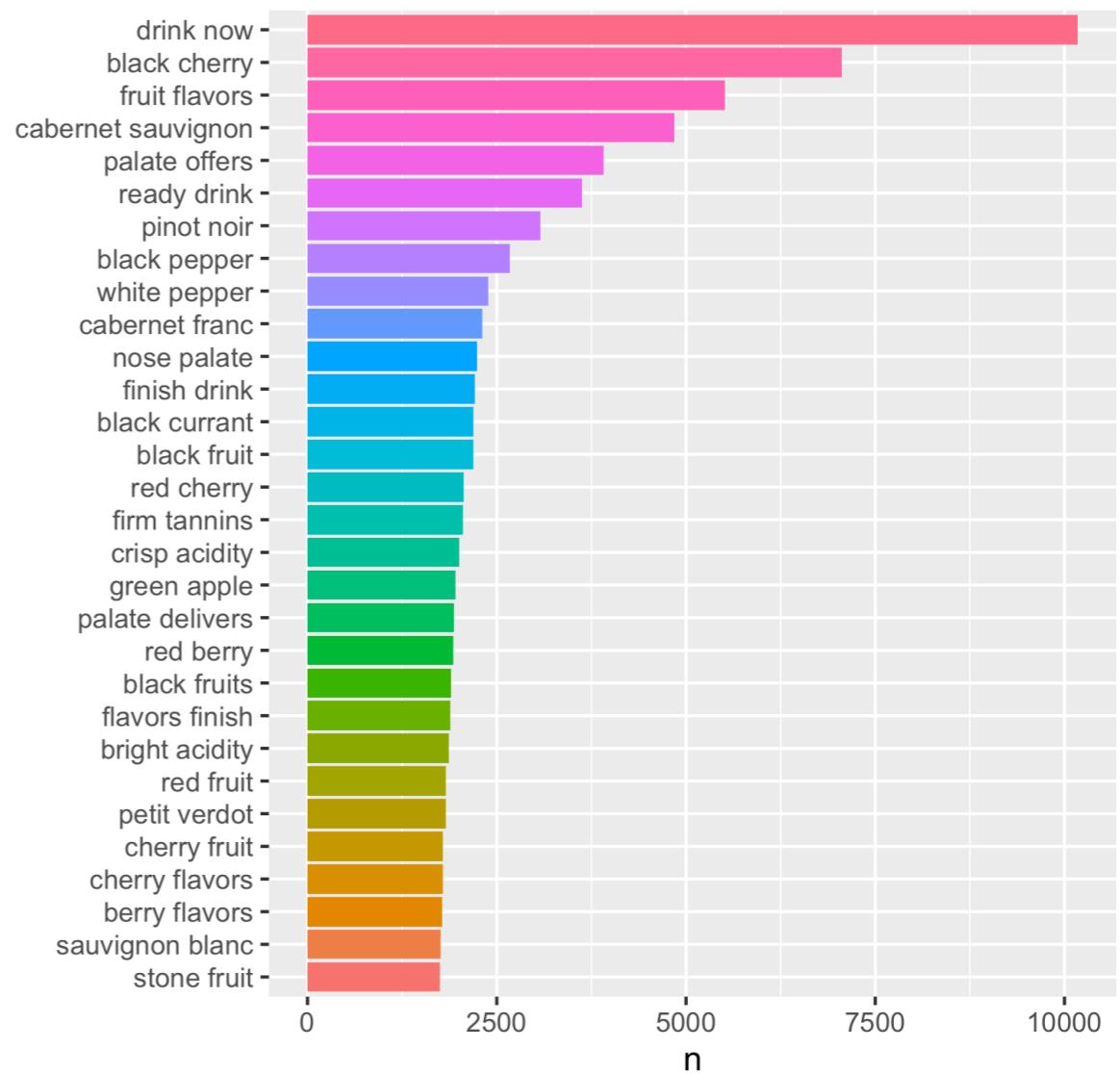


# Expresiones frecuentes

Términos más utilizados



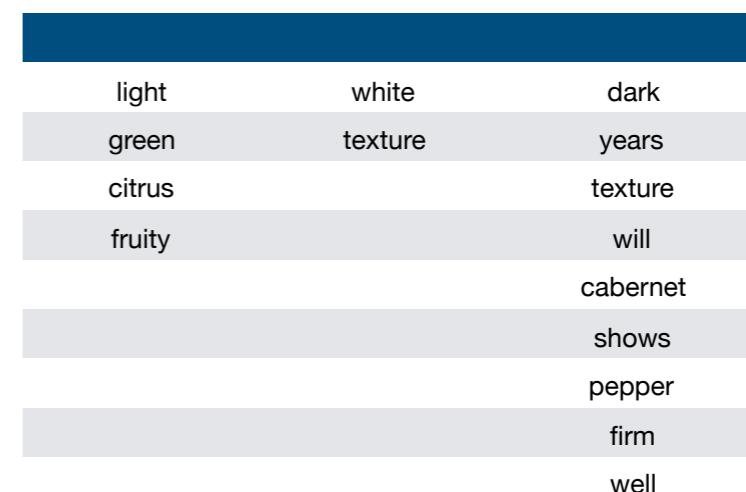
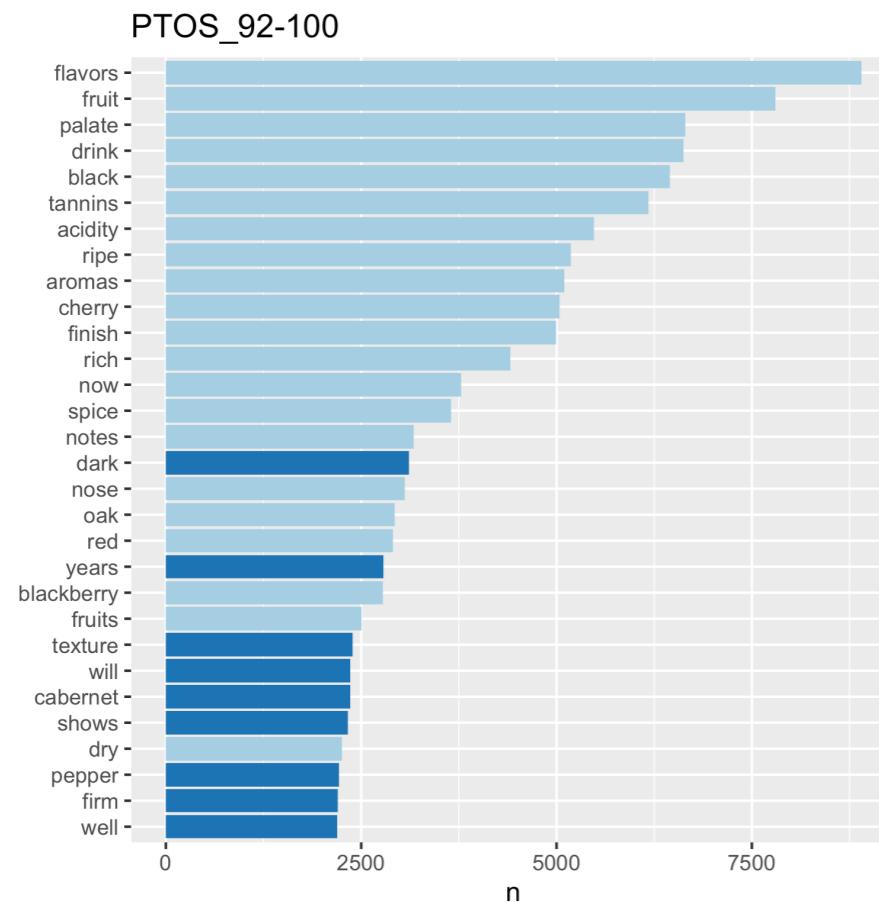
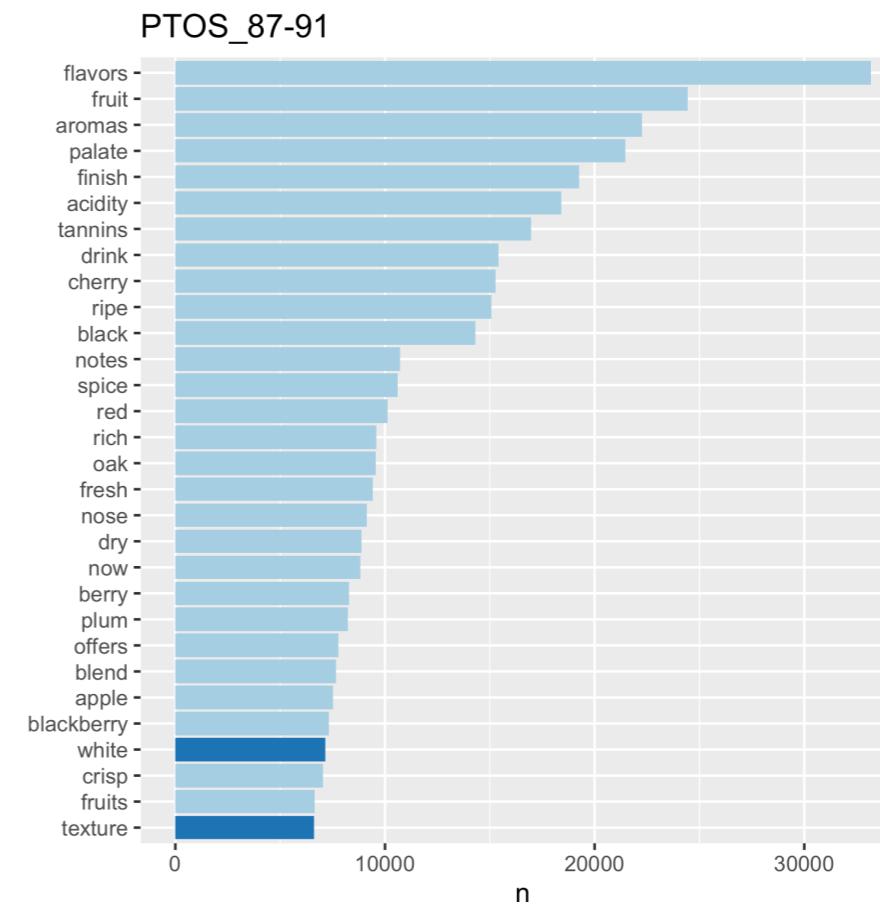
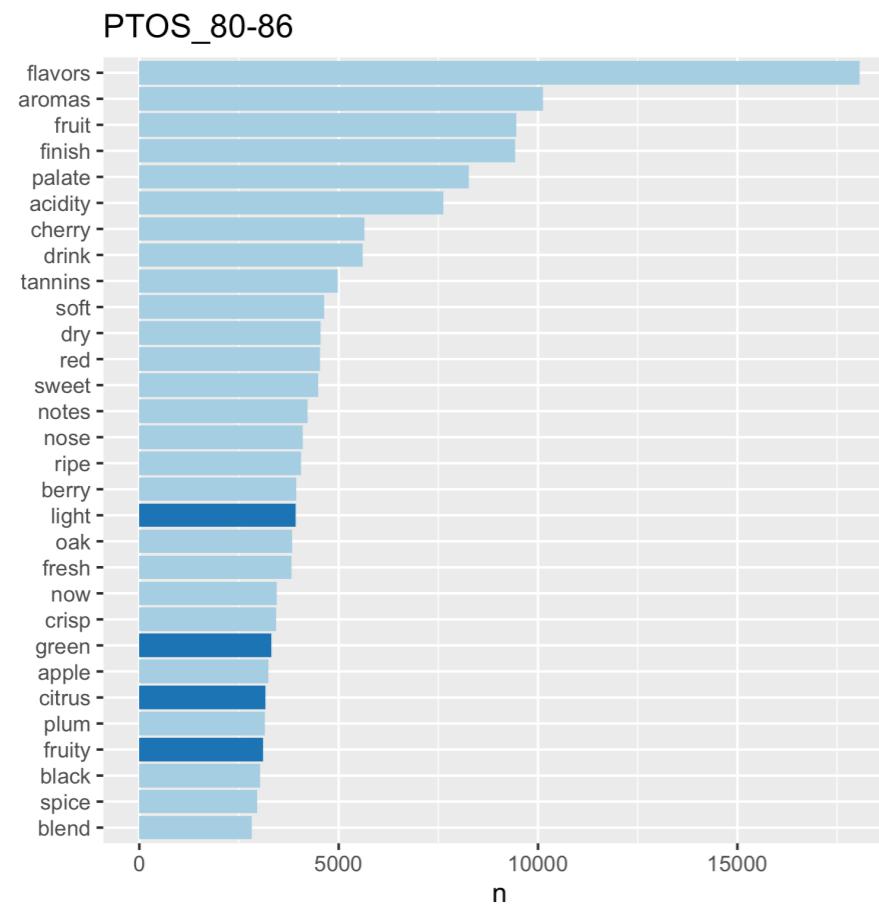
Bigramas

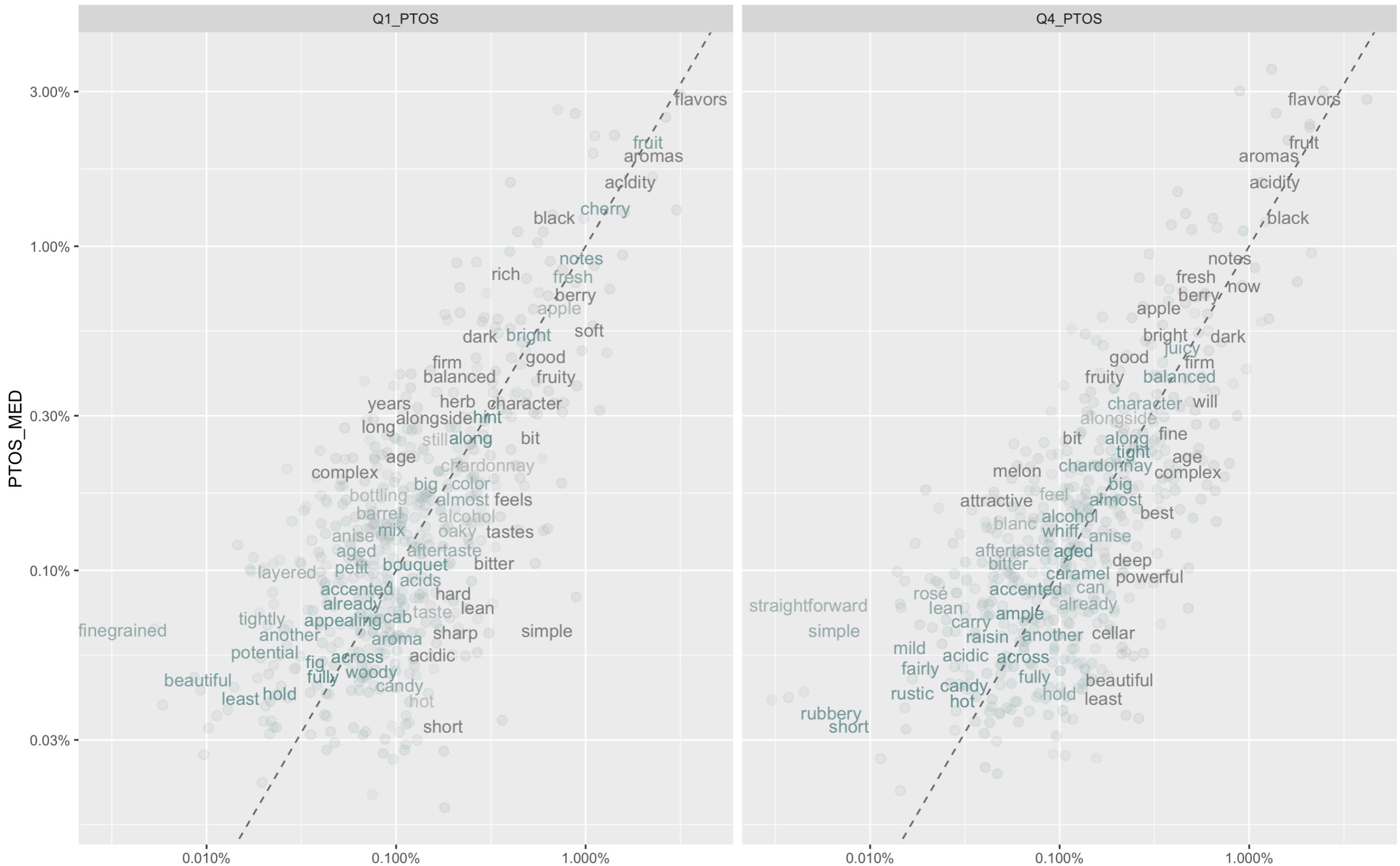


52 concordancias con correlación >= 0.6

20	tannins	firm
25	drink	ready
30	cherrv	cola
31	cherrv	cranberrv
44	oak	new
48	blend	sauvianon
49	apple	bear
50	apple	lemon
51	apple	chardonnnav

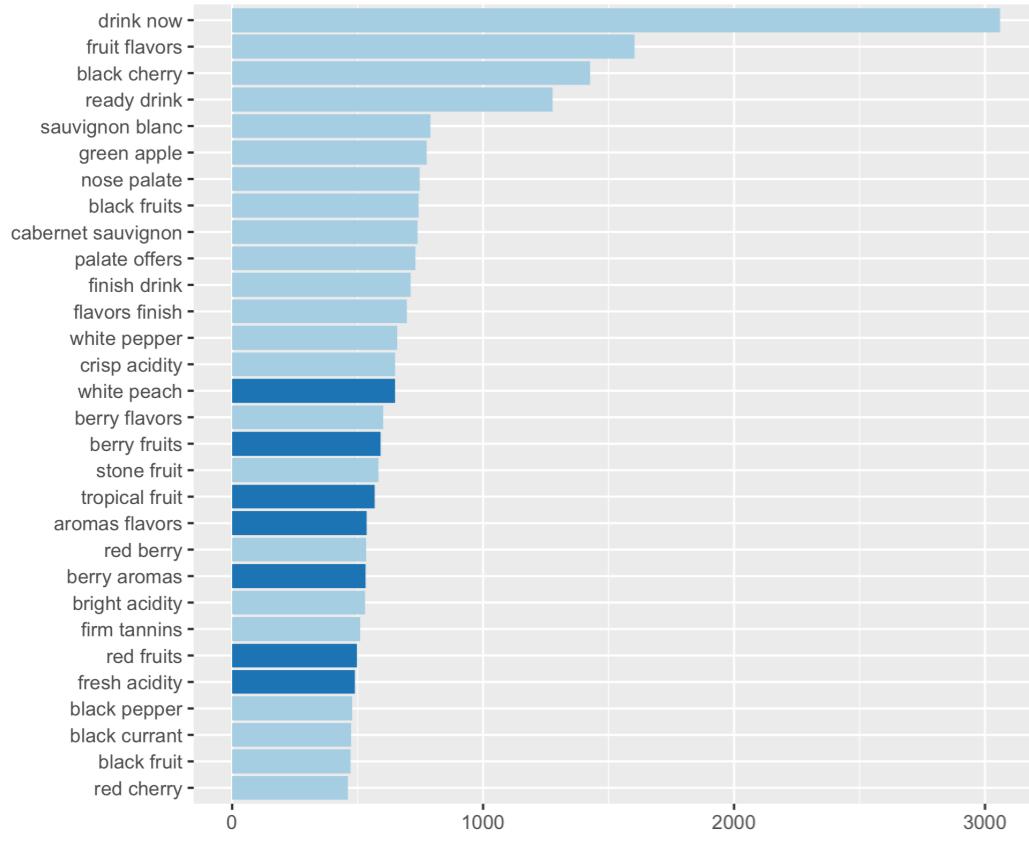
# RANGO\_PTOS - Términos frecuentes



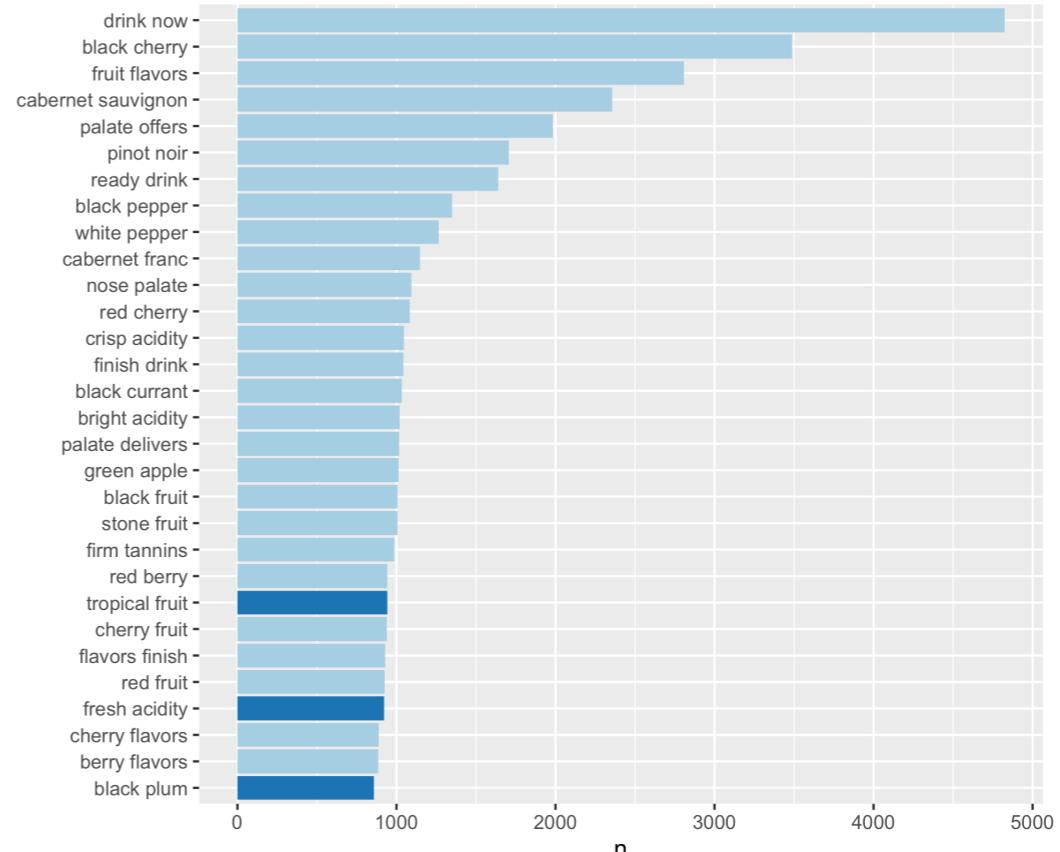


# RANGO\_PRECIO - Bigramas frecuentes

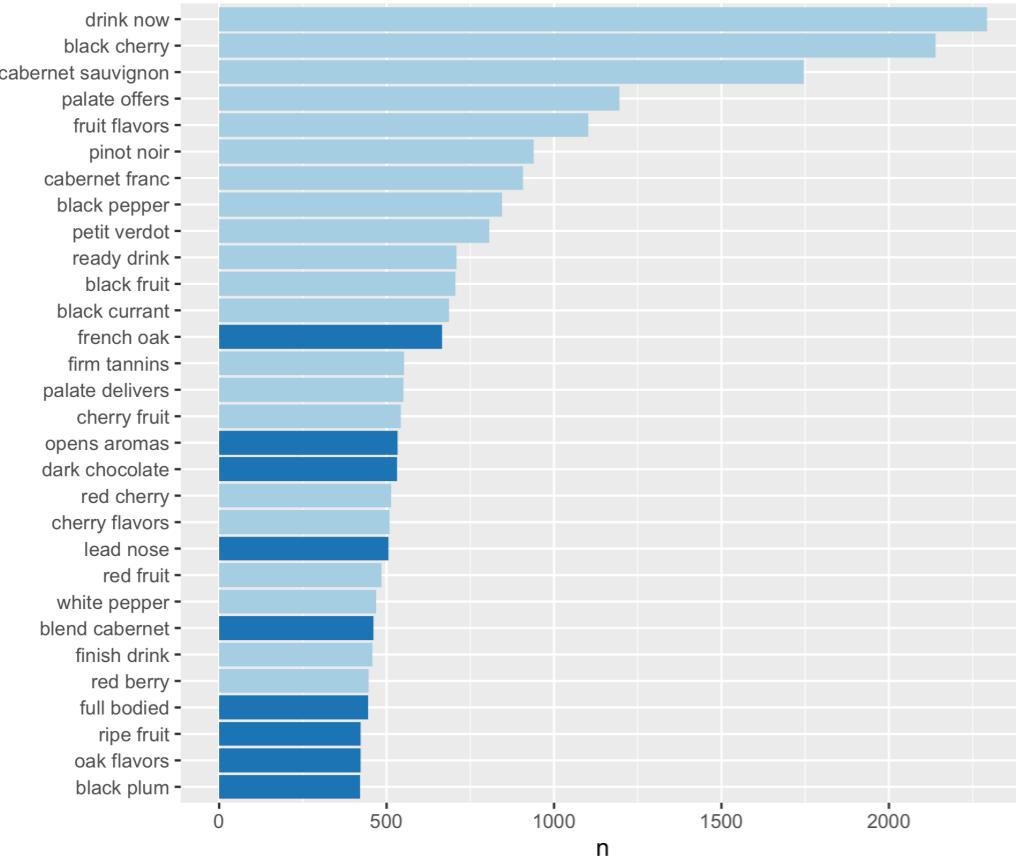
Bigramas\_BARATO



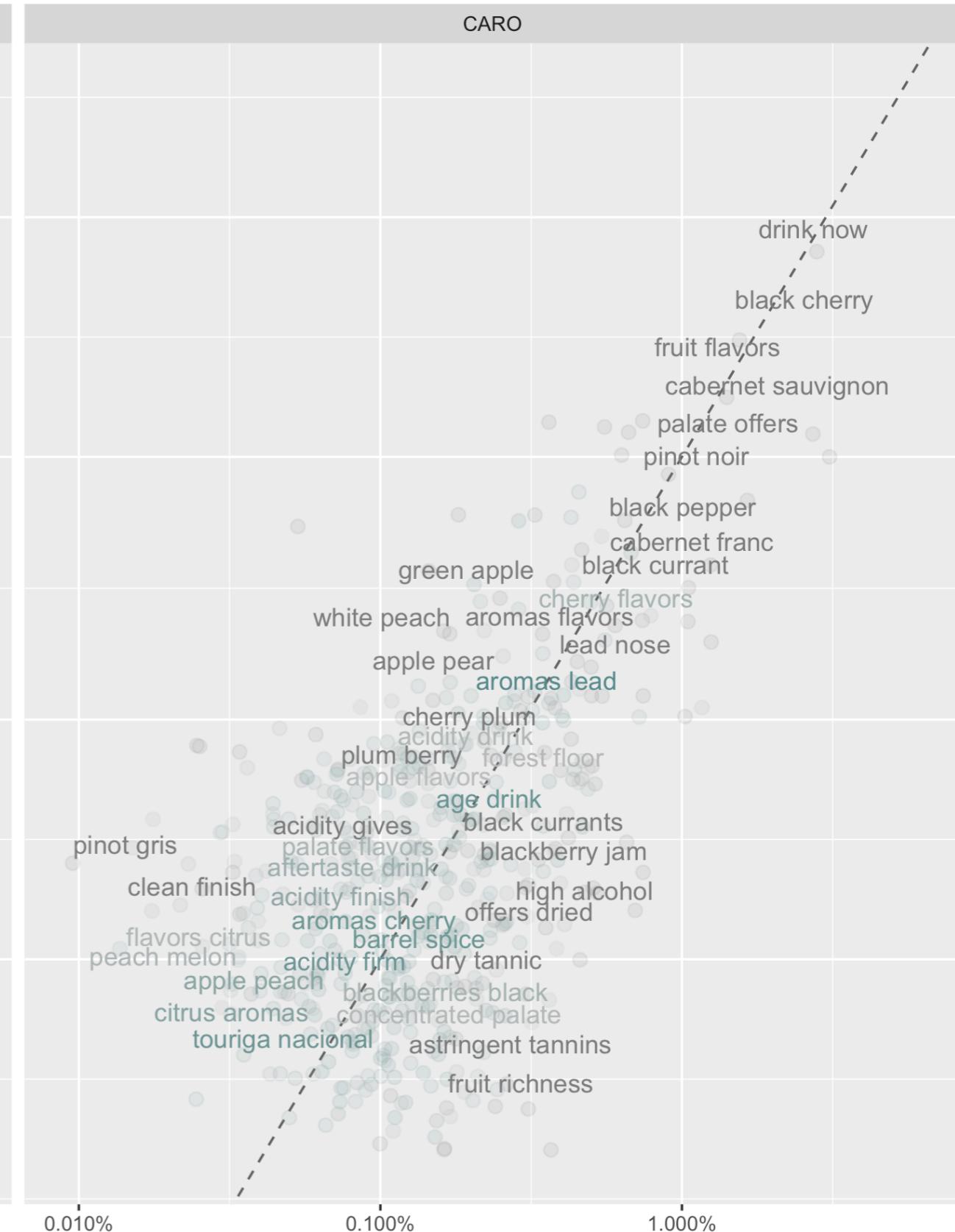
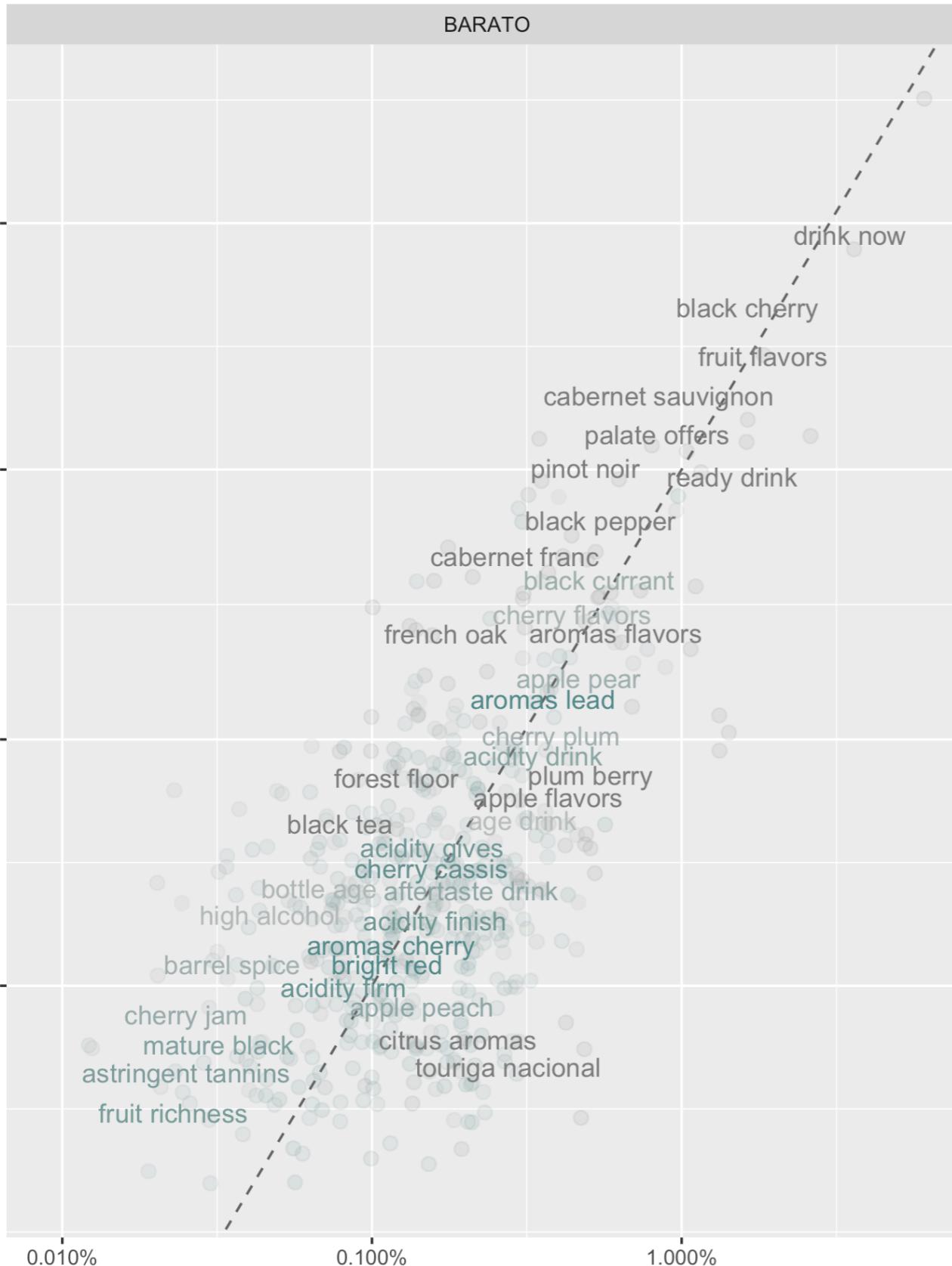
Bigramas\_PRECIO\_MED



Bigramas\_CARO



white peach   tropical fruit   french oak  
berry fruits   fresh acidity   opens aromas  
tropical fruit   black plum   dark chocolate  
aromas flavors   lead nose  
berry aromas   blend cabernet  
red fruits   full bodied  
fresh acidity   ripe fruit  
oak flavors   black plum

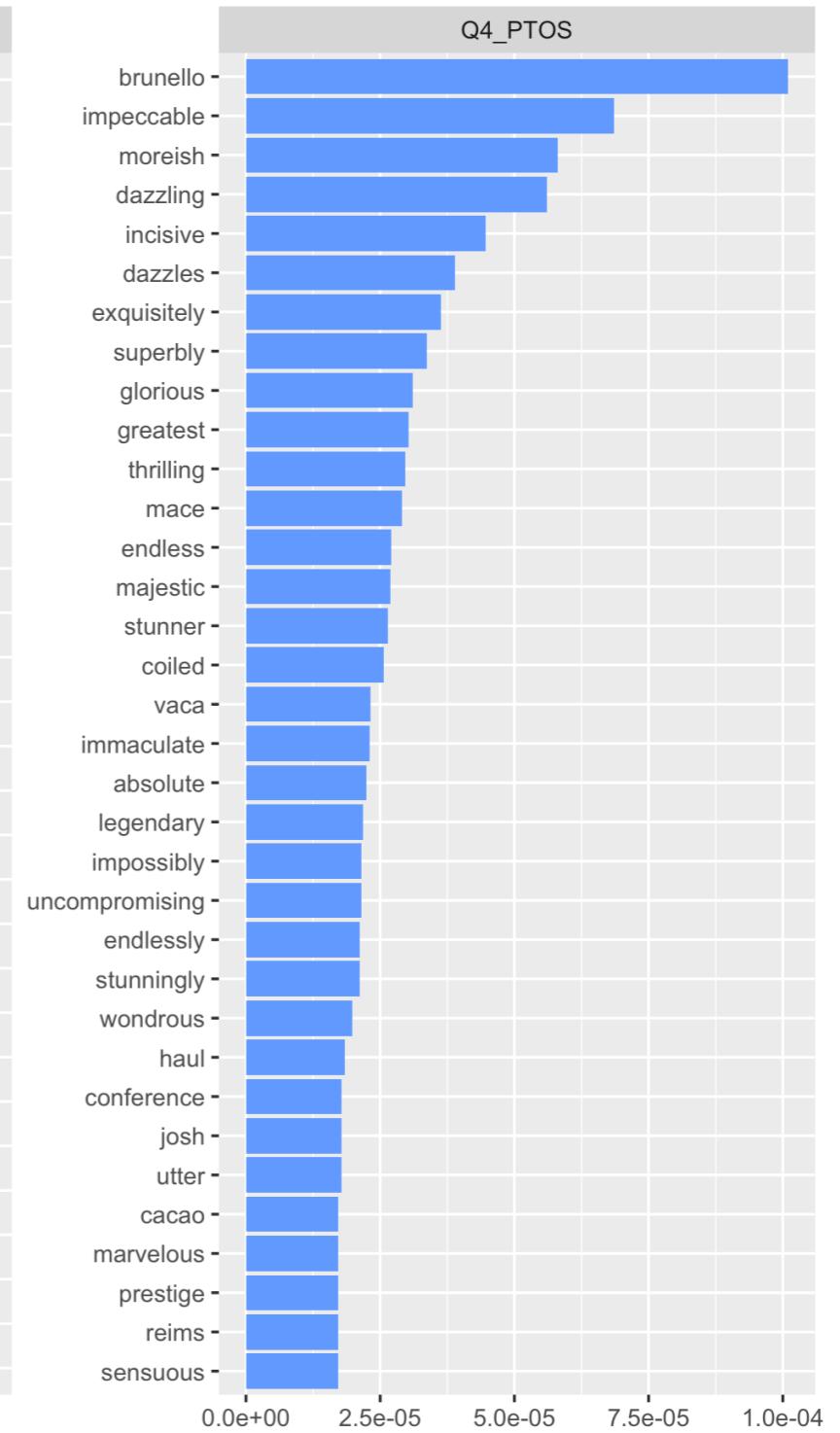
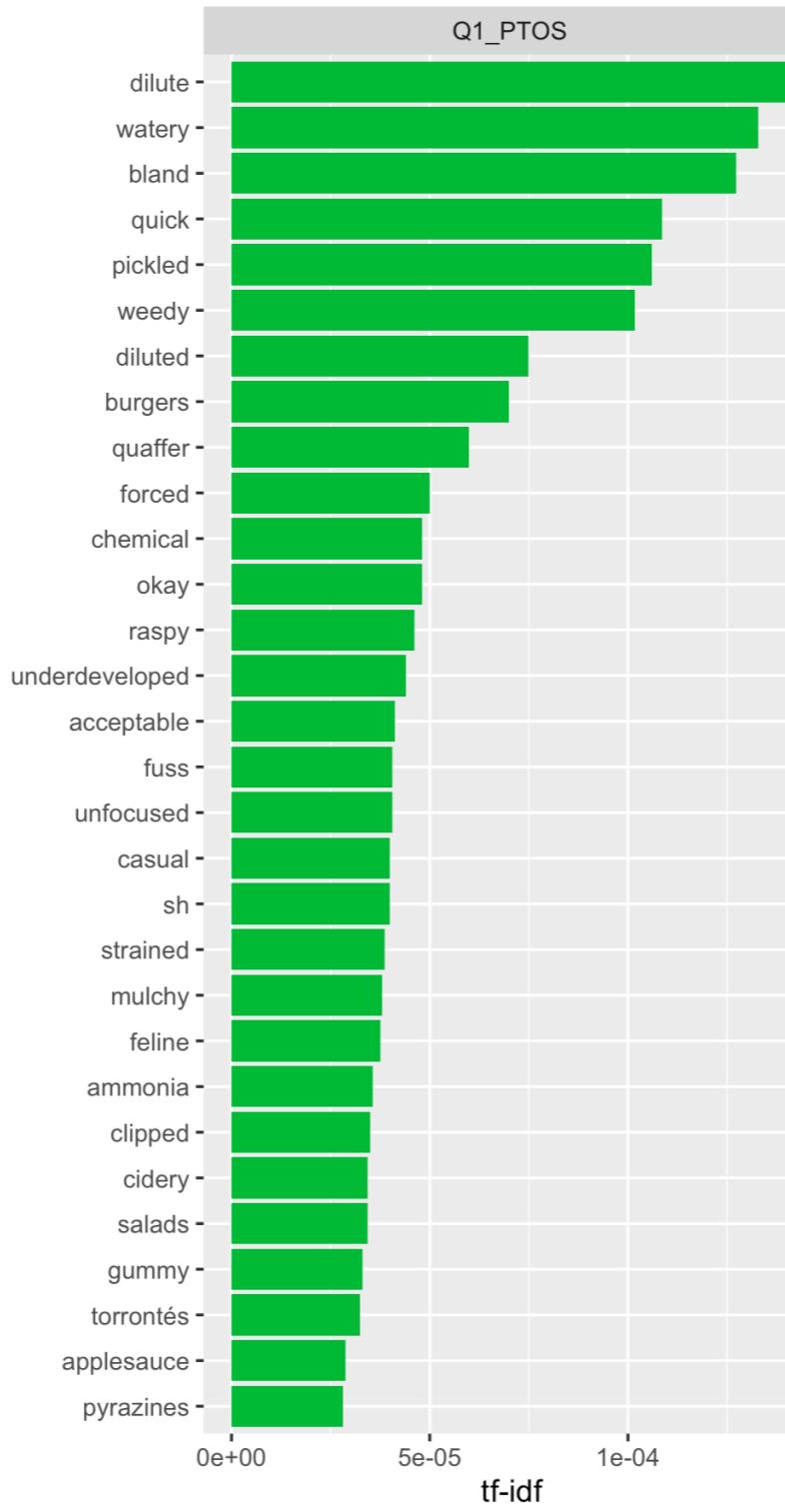
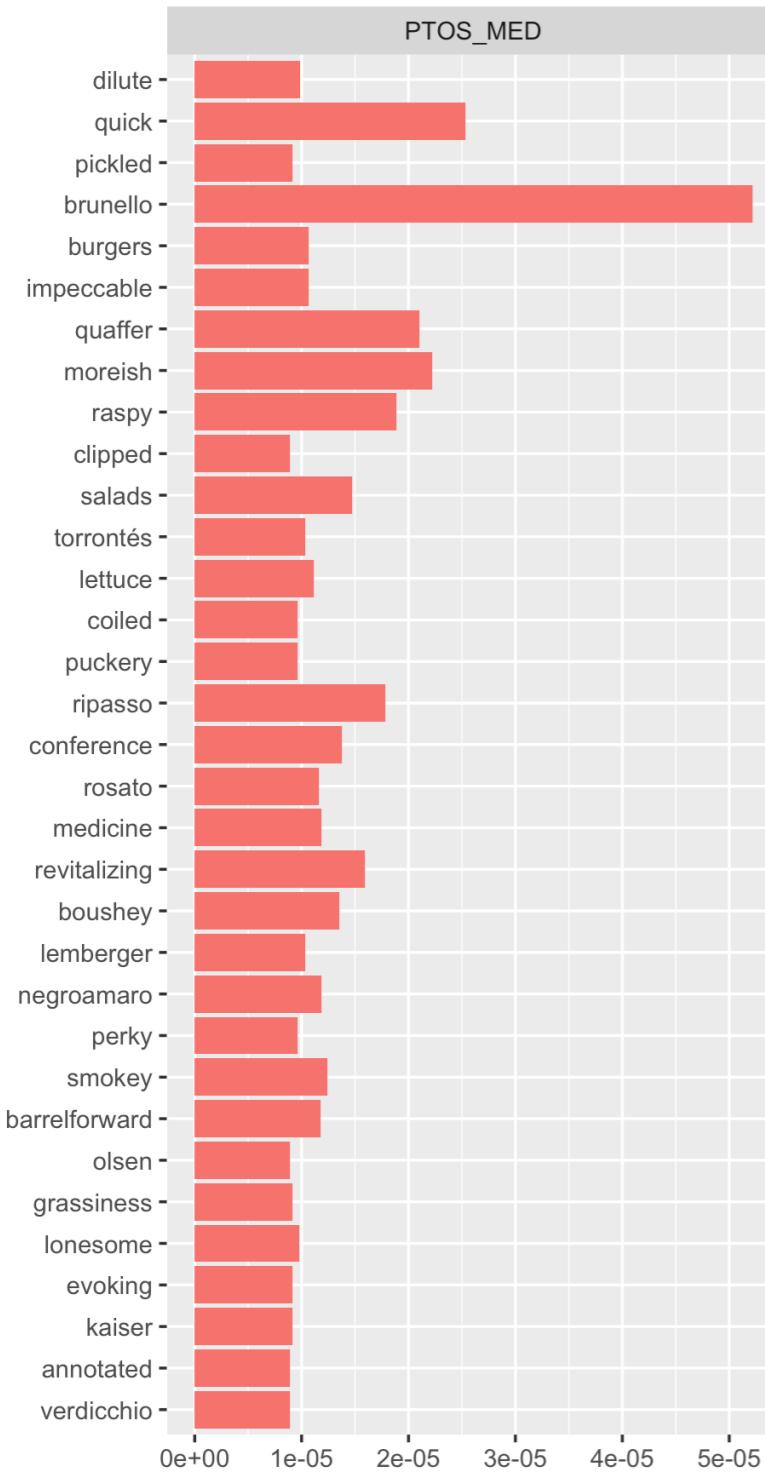


## RANGO PTOS - Términos característicos

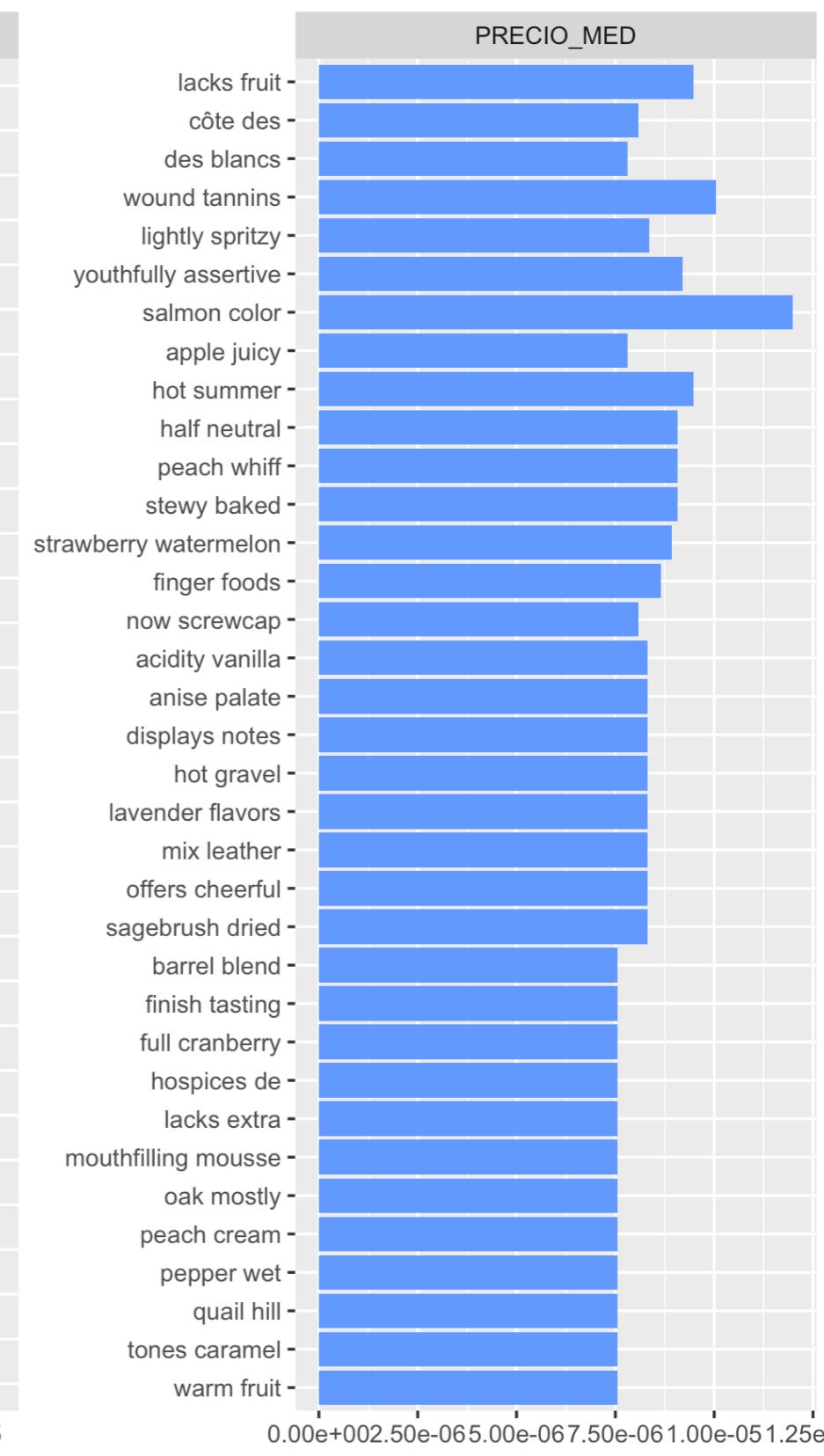
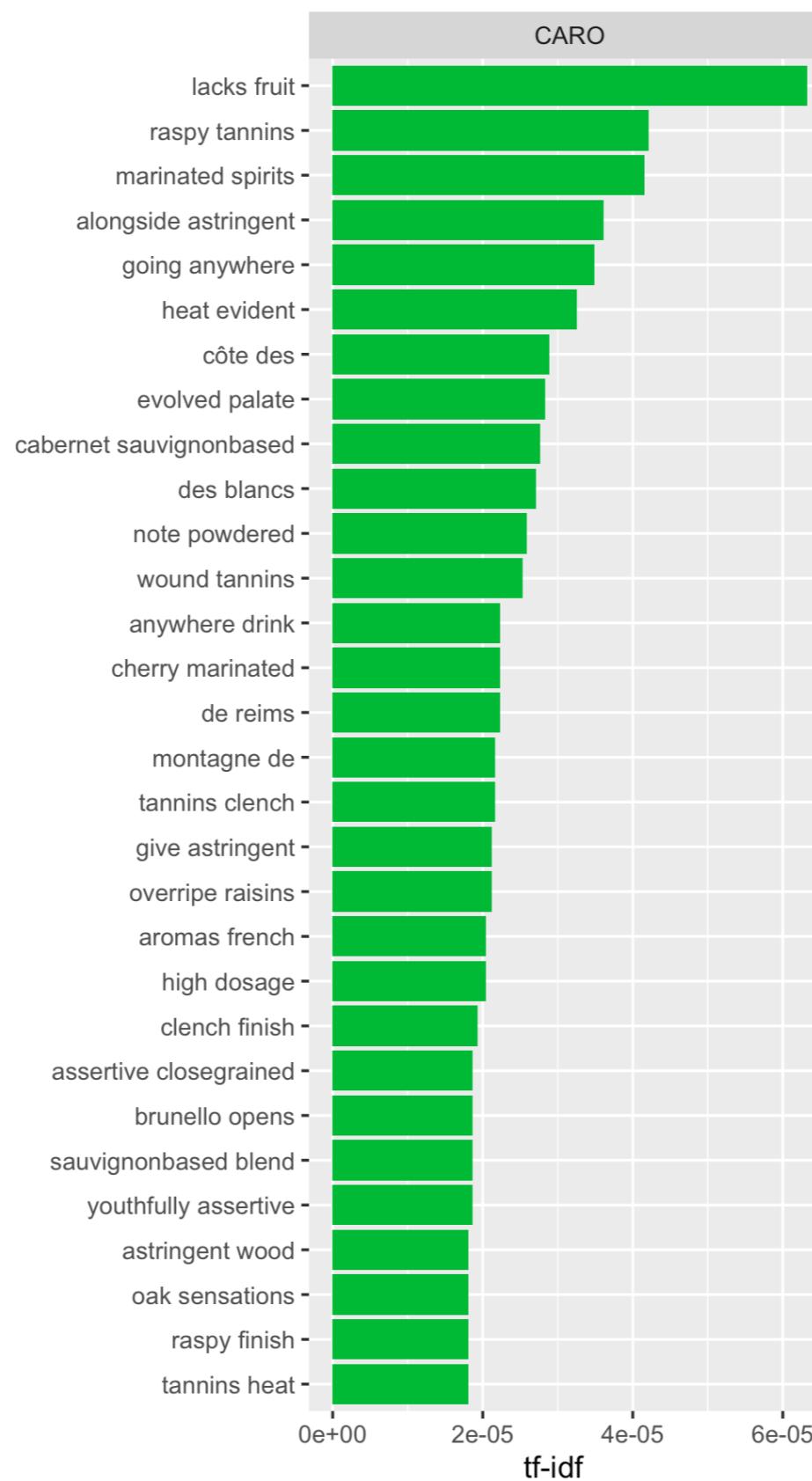
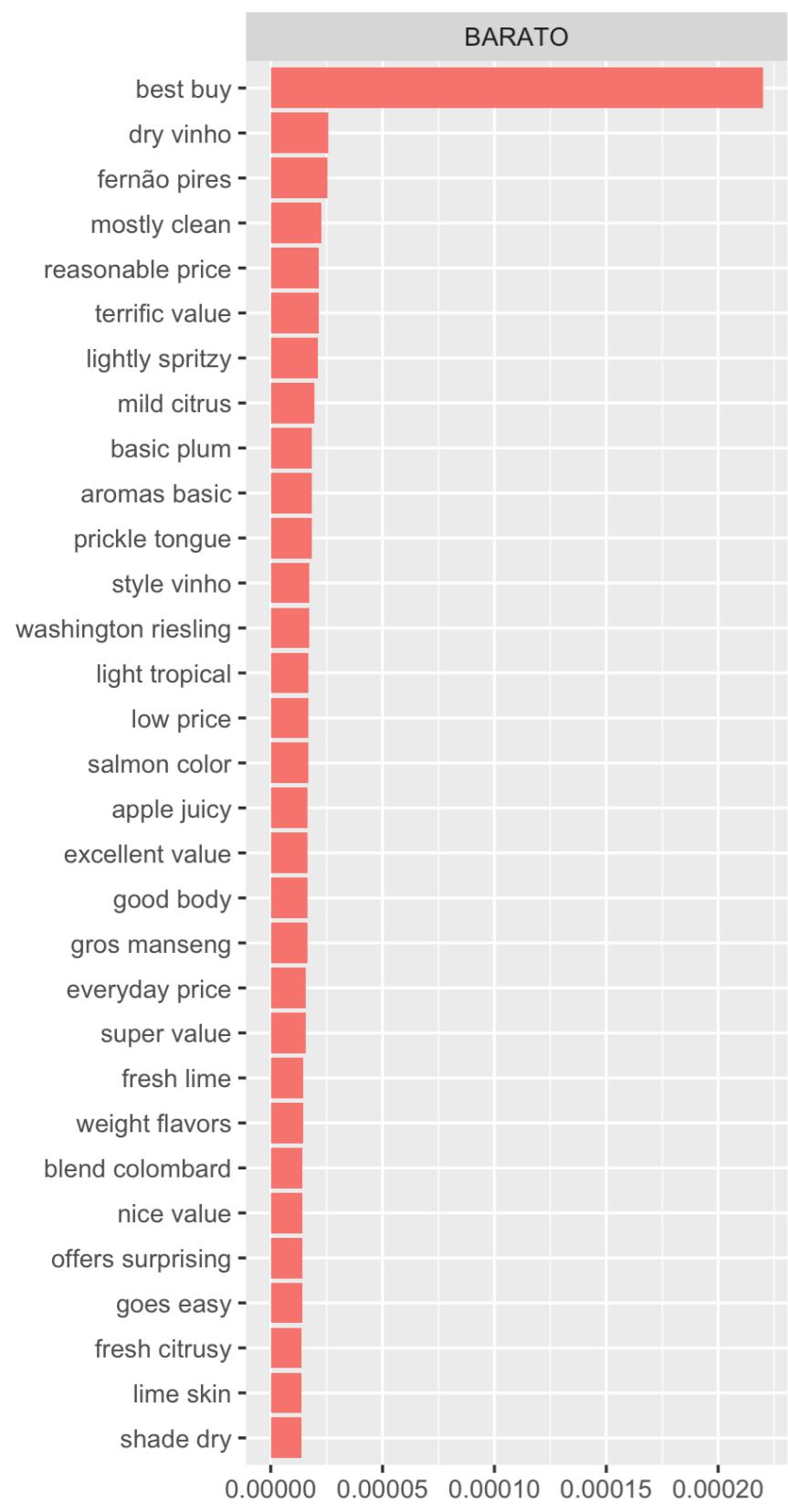
$$\text{tf-idf} = f(\text{termino}) * \ln\left(\frac{n_{\text{documentos}}}{n_{\text{documentos que contienen termino}}}\right)$$

# RANGO PTOS - Términos característicos

$$\text{tf-idf} = f(\text{termino}) * \ln\left(\frac{n_{\text{documentos}}}{n_{\text{documentos que contiene el termino}}}\right)$$



## RANGO PRECIO - Bigramas característicos



# Análisis de Correspondencias

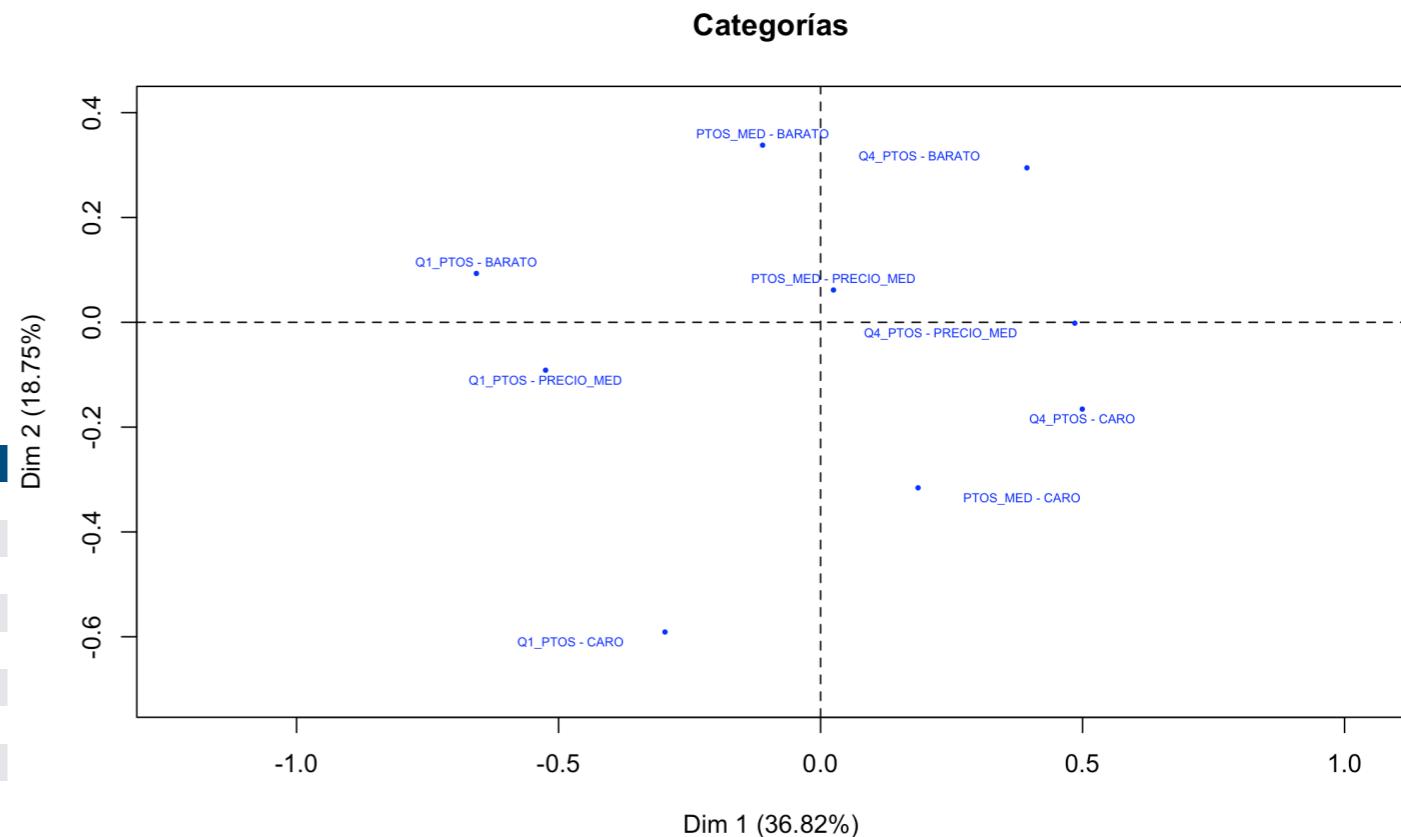
	eigenvalue	% variance	cum. % variance
dim 1	0.11247099	36.822555	36.82256
dim 2	0.05726243	18.747492	55.57005
dim 3	0.03324244	10.883443	66.45349
dim 4	0.02786108	9.121605	75.57509
dim 5	0.02164793	7.087446	82.66254
dim 6	0.01890430	6.189193	88.85173
dim 7	0.01822334	5.966248	94.81798
dim 8	0.01582798	5.182018	100.00000

	Dim_1	Dim_2	Dim_3	Dim_4	Dim_5
PTOS MED - BARATO	1.630321	2.992076e+01	4.02037	0.0229327	19.676582
PTOS MED - CARO	4.058887	2.296593e+01	13.28120	1.3304415	0.124777
PTOS MED - PRECIO MED	0.156103	1.917583e+00	22.69411	5.2032453	7.584196
Q1 PTOS - BARATO	23.525379	9.313385e-01	6.85027	20.6780962	2.276299
Q1 PTOS - CARO	4.105905	3.195225e+01	4.67478	21.1180678	13.223596
Q1 PTOS - PRECIO MED	26.191129	1.559839e+00	5.94940	0.1178163	2.795241
Q4 PTOS - BARATO	7.657455	8.427765e+00	24.94767	15.1666450	3.582012
Q4 PTOS - CARO	10.758486	2.324032e+00	1.05529	35.5989770	26.353060
Q4 PTOS - PRECIO MED	21.916329	5.115547e-04	16.52687	0.7637776	24.384233

# Análisis de Correspondencias

	eigenvalue	% variance	cum. % variance
dim 1	0.11247099	36.822555	36.82256
dim 2	0.05726243	18.747492	55.57005
dim 3	0.03324244	10.883443	66.45349
dim 4	0.02786108	9.121605	75.57509
dim 5	0.02164793	7.087446	82.66254
dim 6	0.01890430	6.189193	88.85173
dim 7	0.01822334	5.966248	94.81798
dim 8	0.01582798	5.182018	100.00000

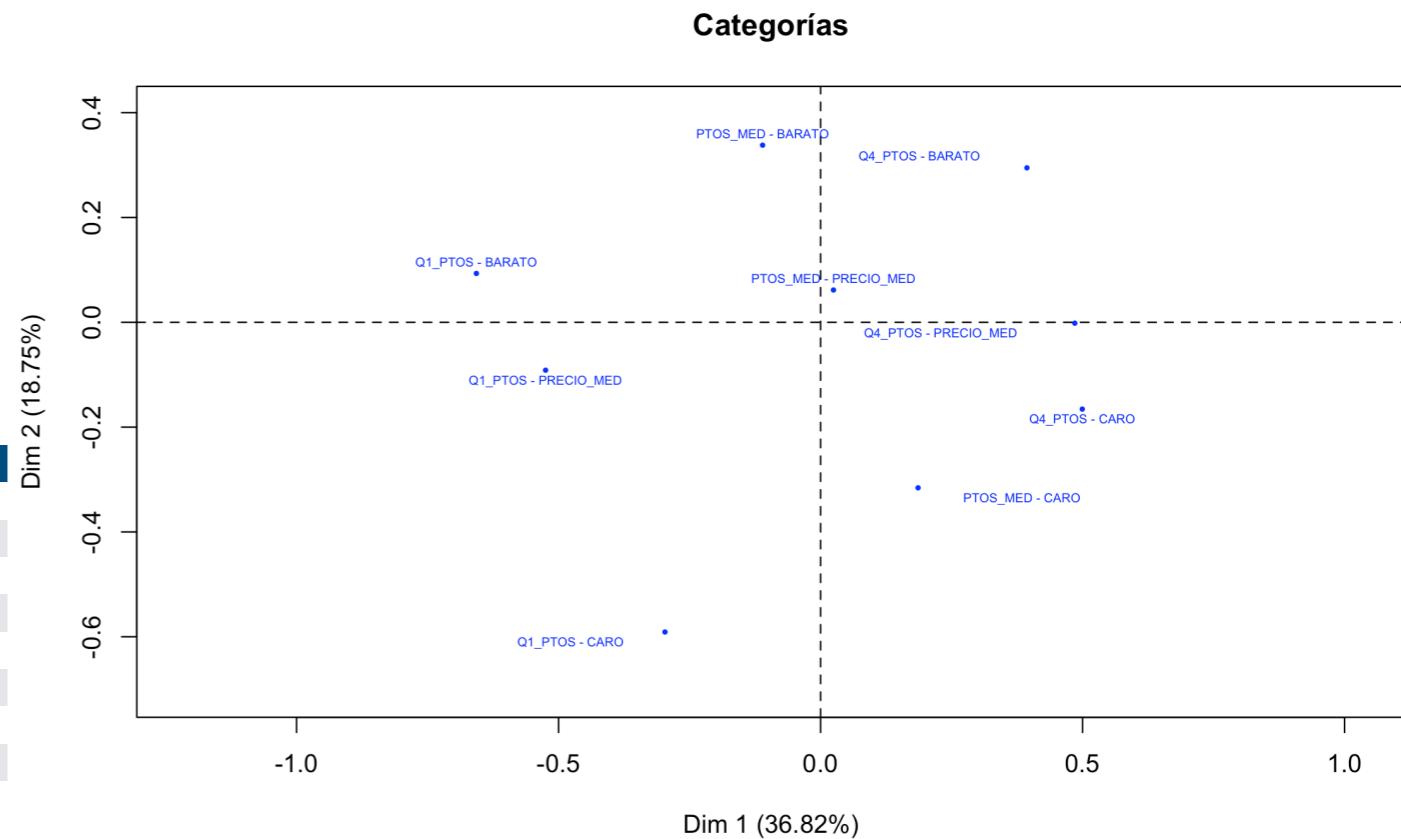
	Dim_1	Dim_2	Dim_3	Dim_4	Dim_5
PTOS MED - BARATO	1.630321	2.992076e+01	4.02037	0.0229327	19.676582
PTOS MED - CARO	4.058887	2.296593e+01	13.28120	1.3304415	0.124777
PTOS MED - PRECIO MED	0.156103	1.917583e+00	22.69411	5.2032453	7.584196
Q1 PTOS - BARATO	23.525379	9.313385e-01	6.85027	20.6780962	2.276299
Q1 PTOS - CARO	4.105905	3.195225e+01	4.67478	21.1180678	13.223596
Q1 PTOS - PRECIO MED	26.191129	1.559839e+00	5.94940	0.1178163	2.795241
Q4 PTOS - BARATO	7.657455	8.427765e+00	24.94767	15.1666450	3.582012
Q4 PTOS - CARO	10.758486	2.324032e+00	1.05529	35.5989770	26.353060
Q4 PTOS - PRECIO MED	21.916329	5.115547e-04	16.52687	0.7637776	24.384233



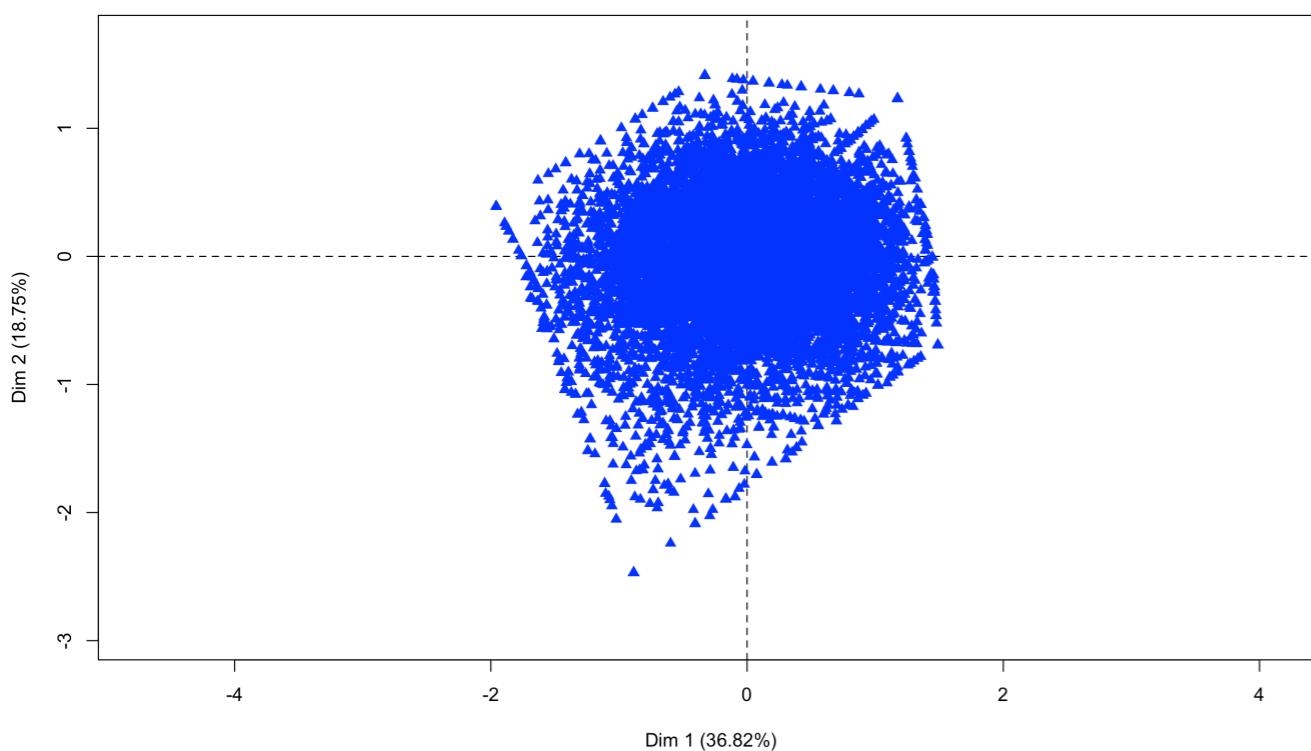
# Análisis de Correspondencias

	eigenvalue	% variance	cum. % variance
dim 1	0.11247099	36.822555	36.82256
dim 2	0.05726243	18.747492	55.57005
dim 3	0.03324244	10.883443	66.45349
dim 4	0.02786108	9.121605	75.57509
dim 5	0.02164793	7.087446	82.66254
dim 6	0.01890430	6.189193	88.85173
dim 7	0.01822334	5.966248	94.81798
dim 8	0.01582798	5.182018	100.00000

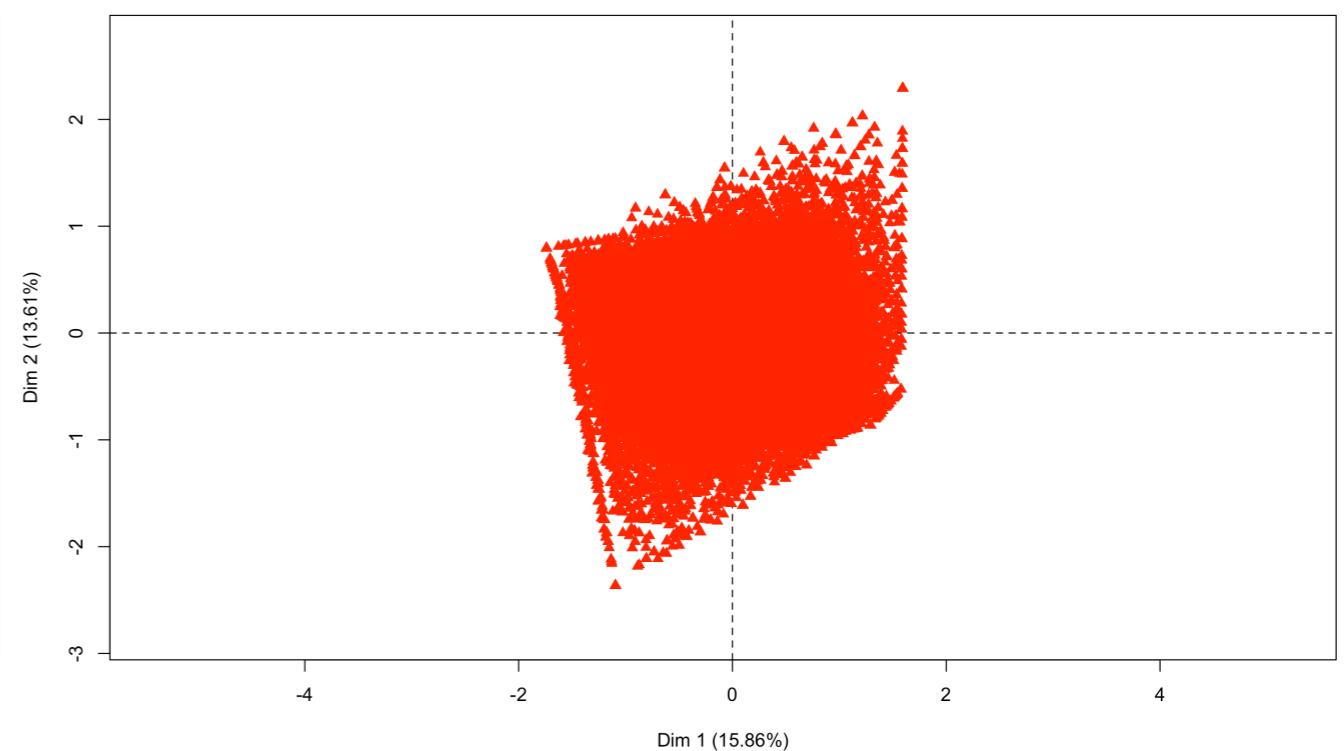
	Dim_1	Dim_2	Dim_3	Dim_4	Dim_5
PTOS MED - BARATO	1.630321	2.992076e+01	4.02037	0.0229327	19.676582
PTOS MED - CARO	4.058887	2.296593e+01	13.28120	1.3304415	0.124777
PTOS MED - PRECIO MED	0.156103	1.917583e+00	22.69411	5.2032453	7.584196
Q1 PTOS - BARATO	23.525379	9.313385e-01	6.85027	20.6780962	2.276299
Q1 PTOS - CARO	4.105905	3.195225e+01	4.67478	21.1180678	13.223596
Q1 PTOS - PRECIO MED	26.191129	1.559839e+00	5.94940	0.1178163	2.795241
Q4 PTOS - BARATO	7.657455	8.427765e+00	24.94767	15.1666450	3.582012
Q4 PTOS - CARO	10.758486	2.324032e+00	1.05529	35.5989770	26.353060
Q4 PTOS - PRECIO MED	21.916329	5.115547e-04	16.52687	0.7637776	24.384233



TÉRMINOS

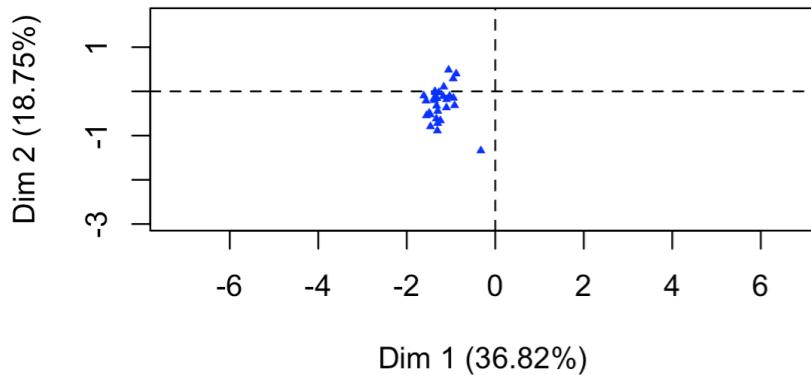


BIGRAMAS

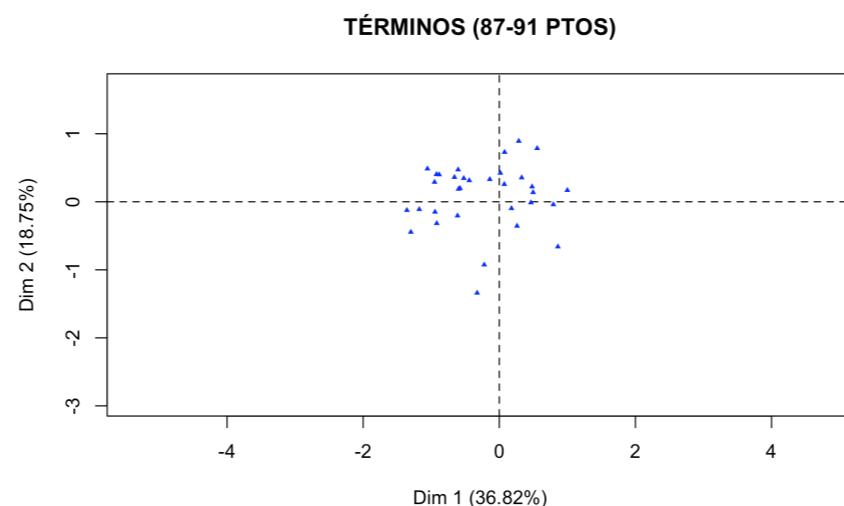


# Análisis de Correspondencias

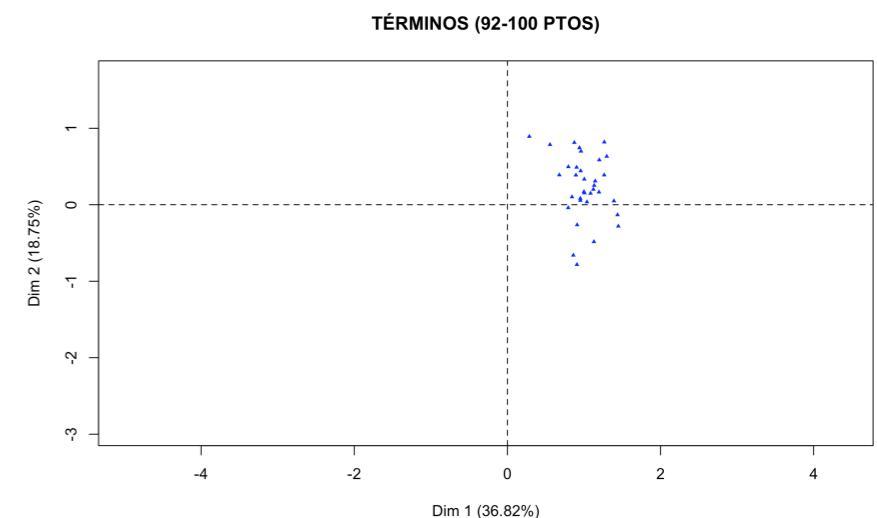
TÉRMINOS (80-86 PTOS)



TÉRMINOS (87-91 PTOS)

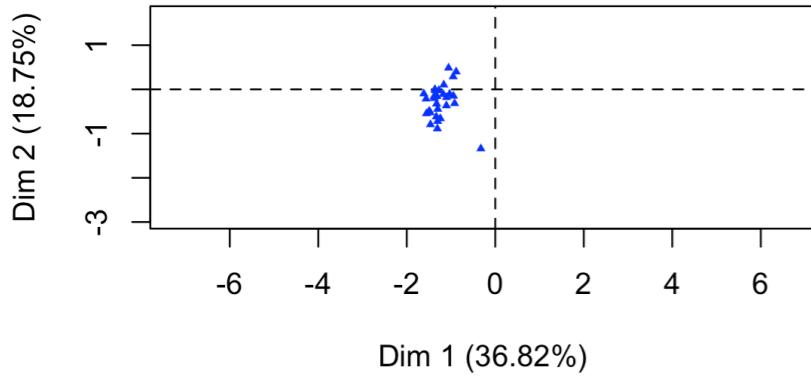


TÉRMINOS (92-100 PTOS)

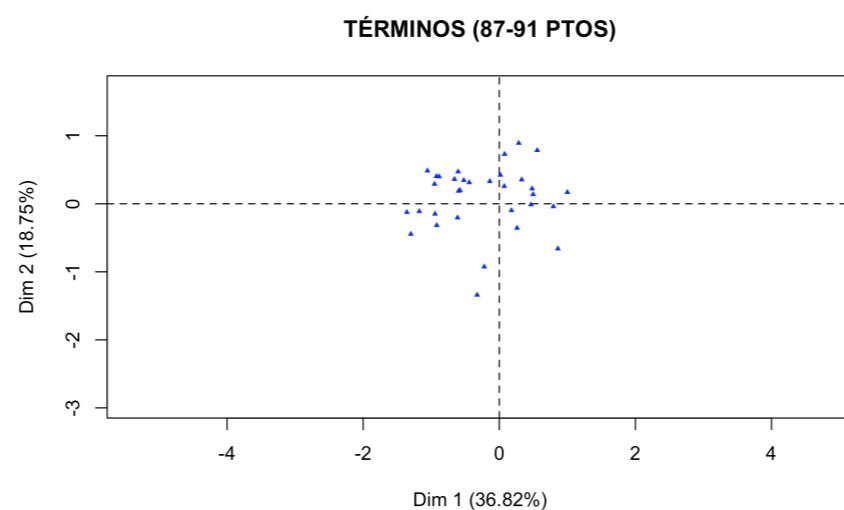


# Análisis de Correspondencias

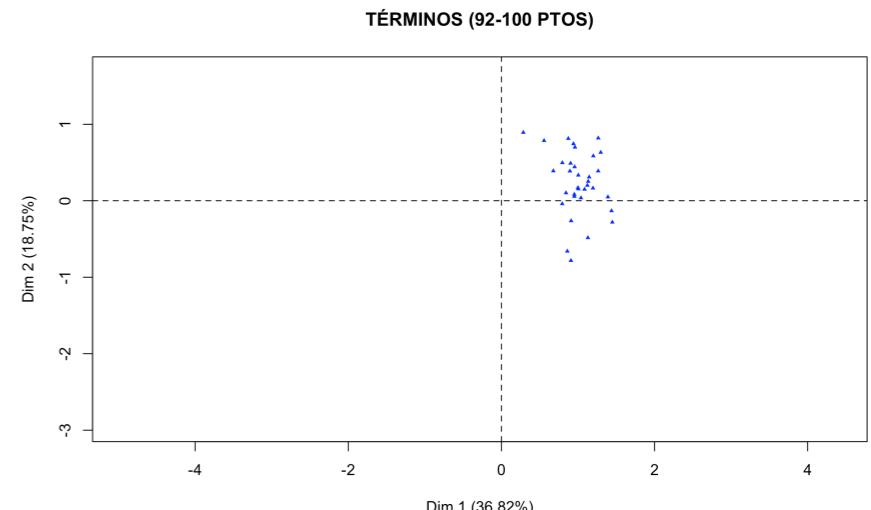
TÉRMINOS (80-86 PTOS)



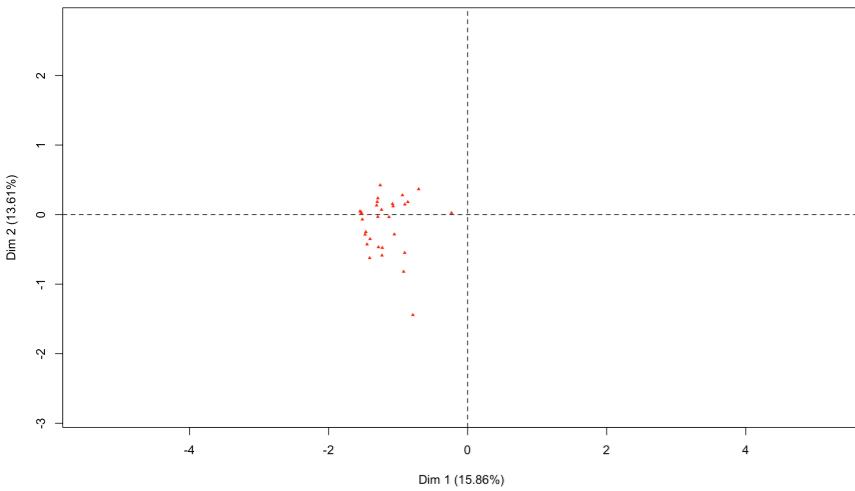
TÉRMINOS (87-91 PTOS)



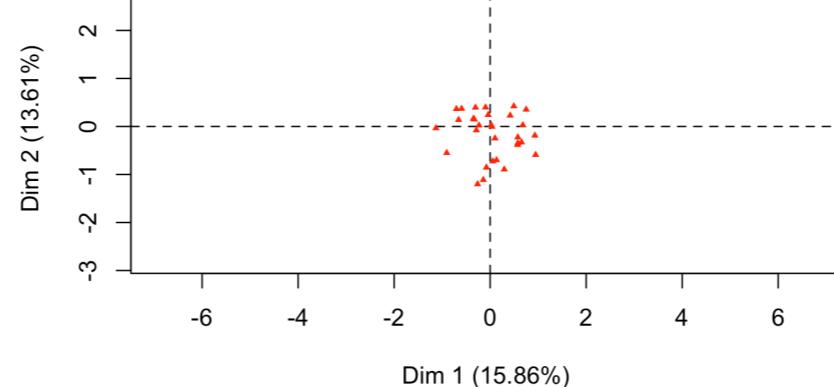
TÉRMINOS (92-100 PTOS)



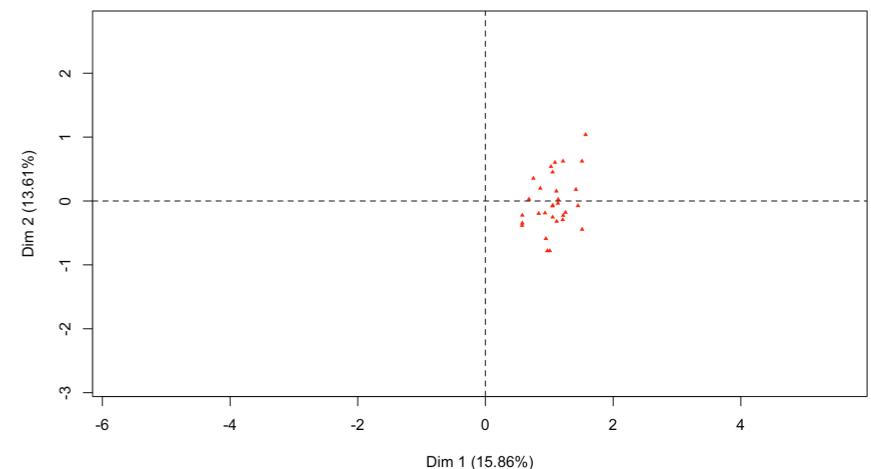
BIGRAMAS (80-86 PTOS)



BIGRAMAS (87-91 PTOS)

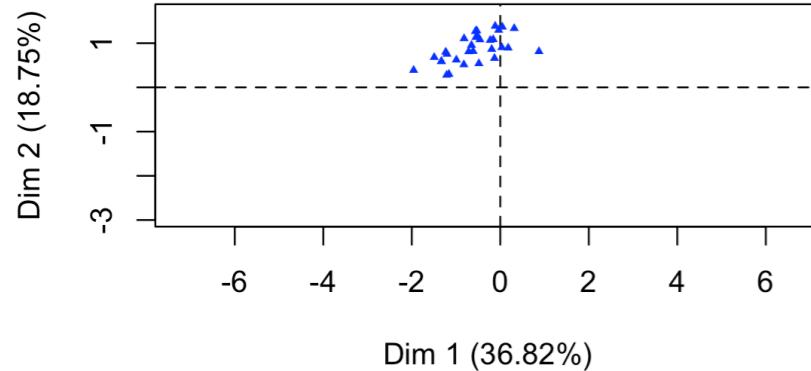


BIGRAMAS (92-100 PUNTOS)

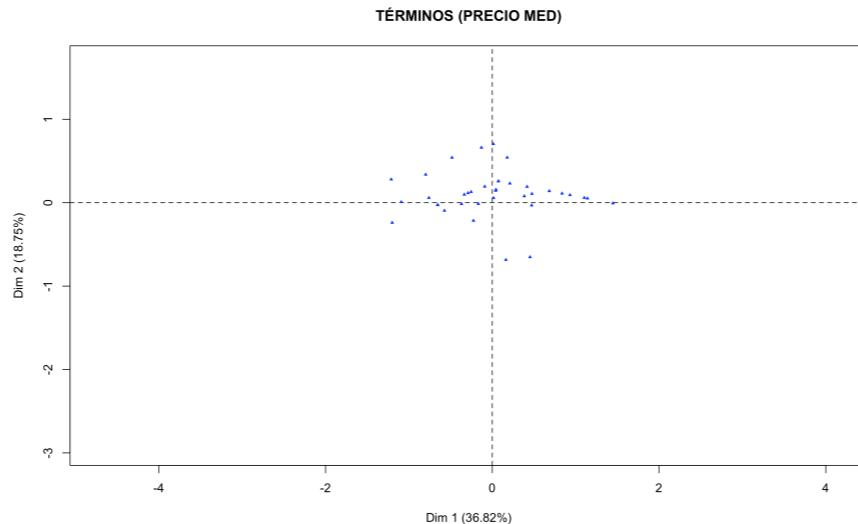


# Análisis de Correspondencias

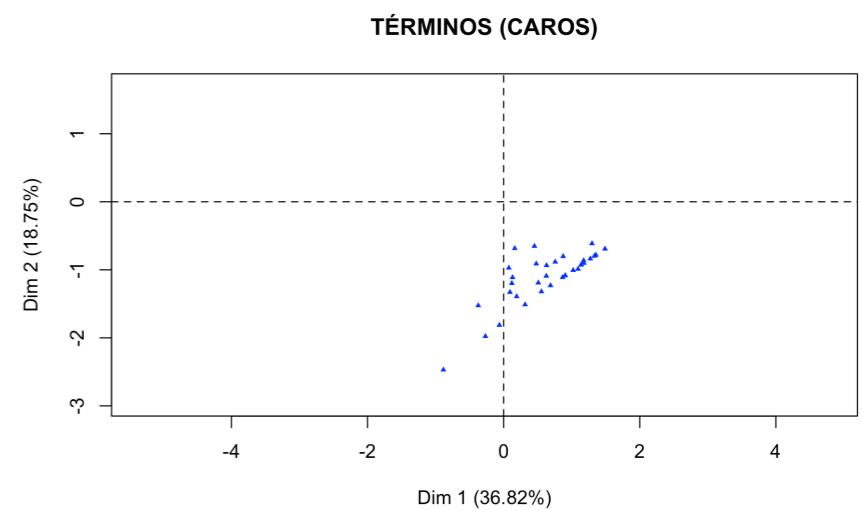
TÉRMINOS (BARATOS)



TÉRMINOS (PRECIO MED)

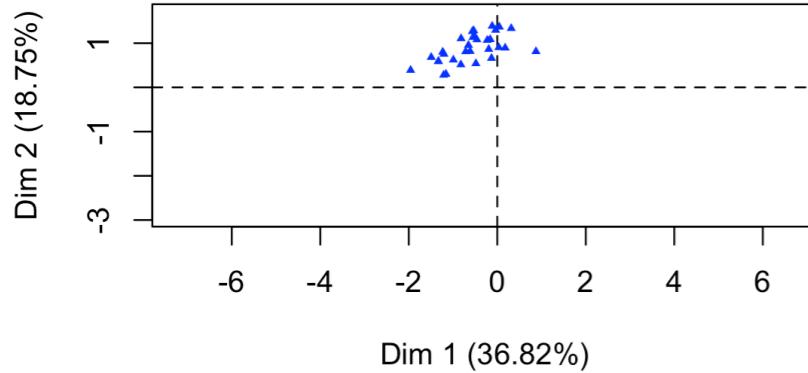


TÉRMINOS (CAROS)

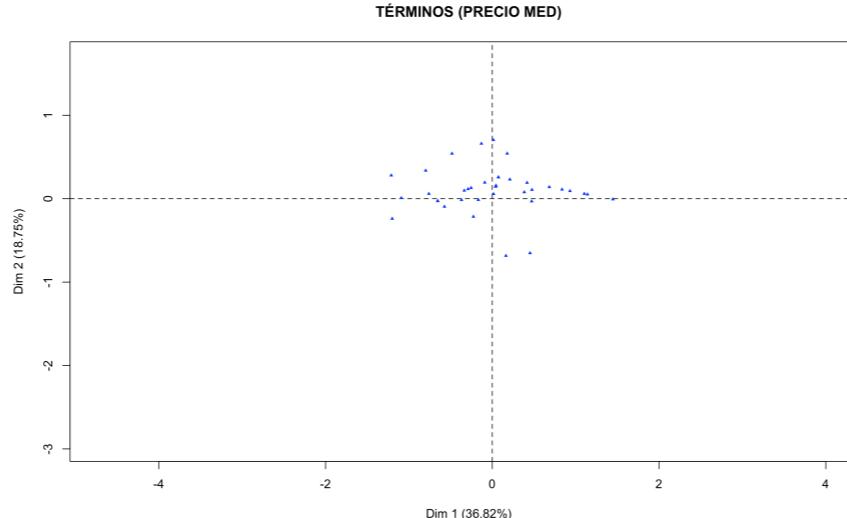


# Análisis de Correspondencias

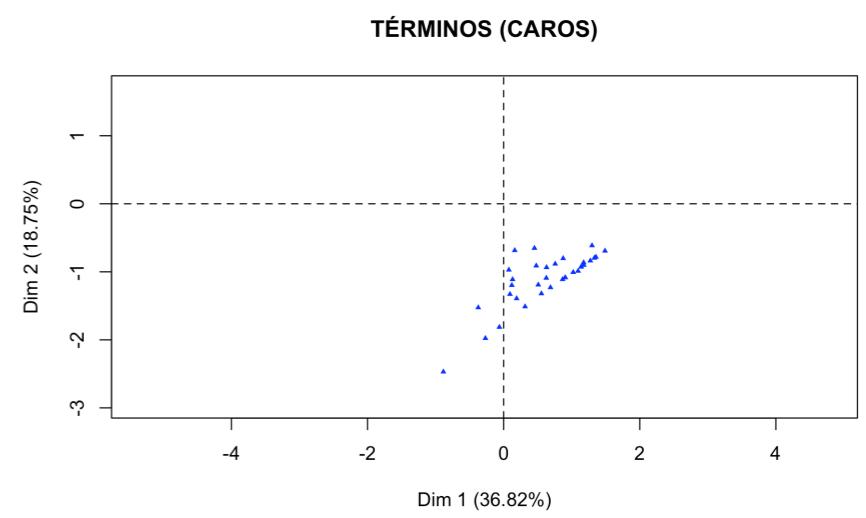
TÉRMINOS (BARATOS)



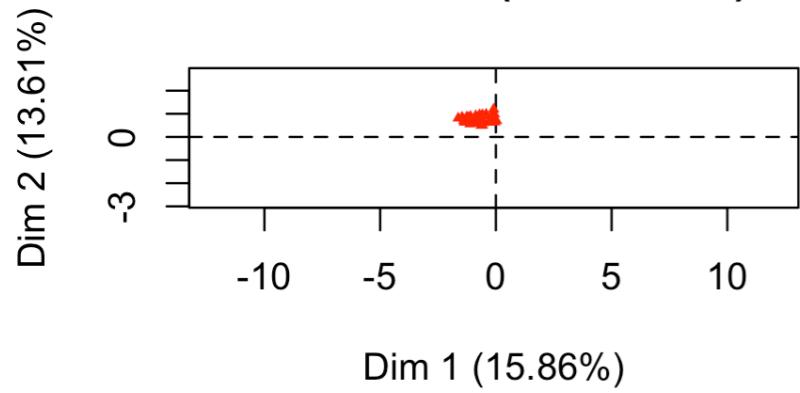
TÉRMINOS (PRECIO MED)



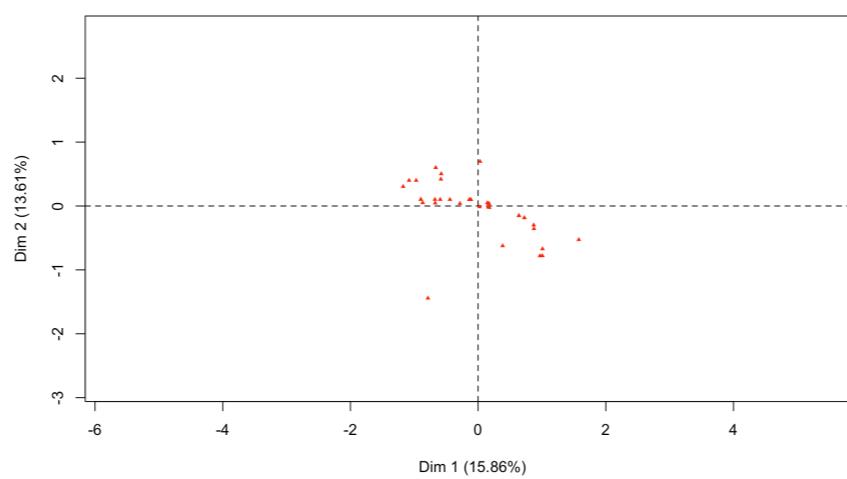
TÉRMINOS (CAROS)



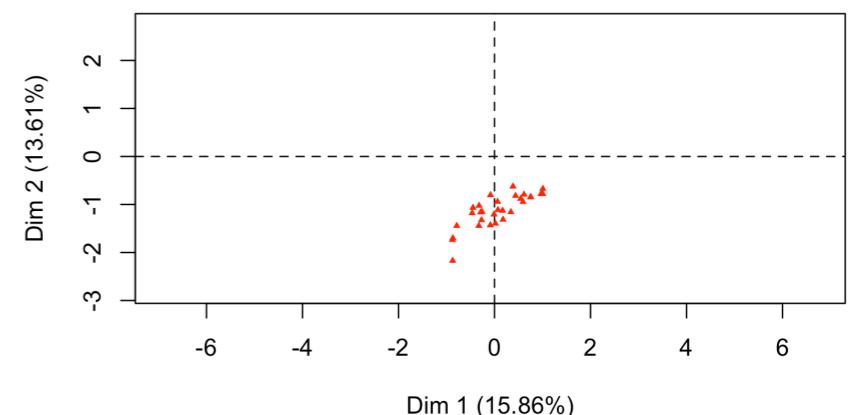
BIGRAMAS (BARATOS)



BIGRAMAS (PRECIO MED)



BIGRAMAS (CAROS)



## Conclusiones

- **Medidas estadísticas y representaciones gráficas más utilizados.**
- **Identificación de expresiones características por categorías de documentos**
- **Análisis de correspondencias**

## Bibliografía

- **Bécue-Bertaut, Mónica:**
  - “*Minería de textos. Aplicación a preguntas abiertas en encuestas.*”
- **Fernández Aguirre, Karmele:**
  - “*Análisis textual: Generación y aplicaciones.*”
- **Terrádez Gurrea, Marcial:**
  - “*Frecuencias Léxicas del Español coloquial: Análisis Cuantitativo y Cualitativo.*”
- **Peña Sánchez de Rivera, Daniel:**
  - “*Estadística. Modelos y Métodos. (1. Fundamentos)*”
  - “*Análisis de Datos Multivariantes.*”

<https://github.com/jdavidfdez/TFG>

***¡ Muchas gracias por su atención !***