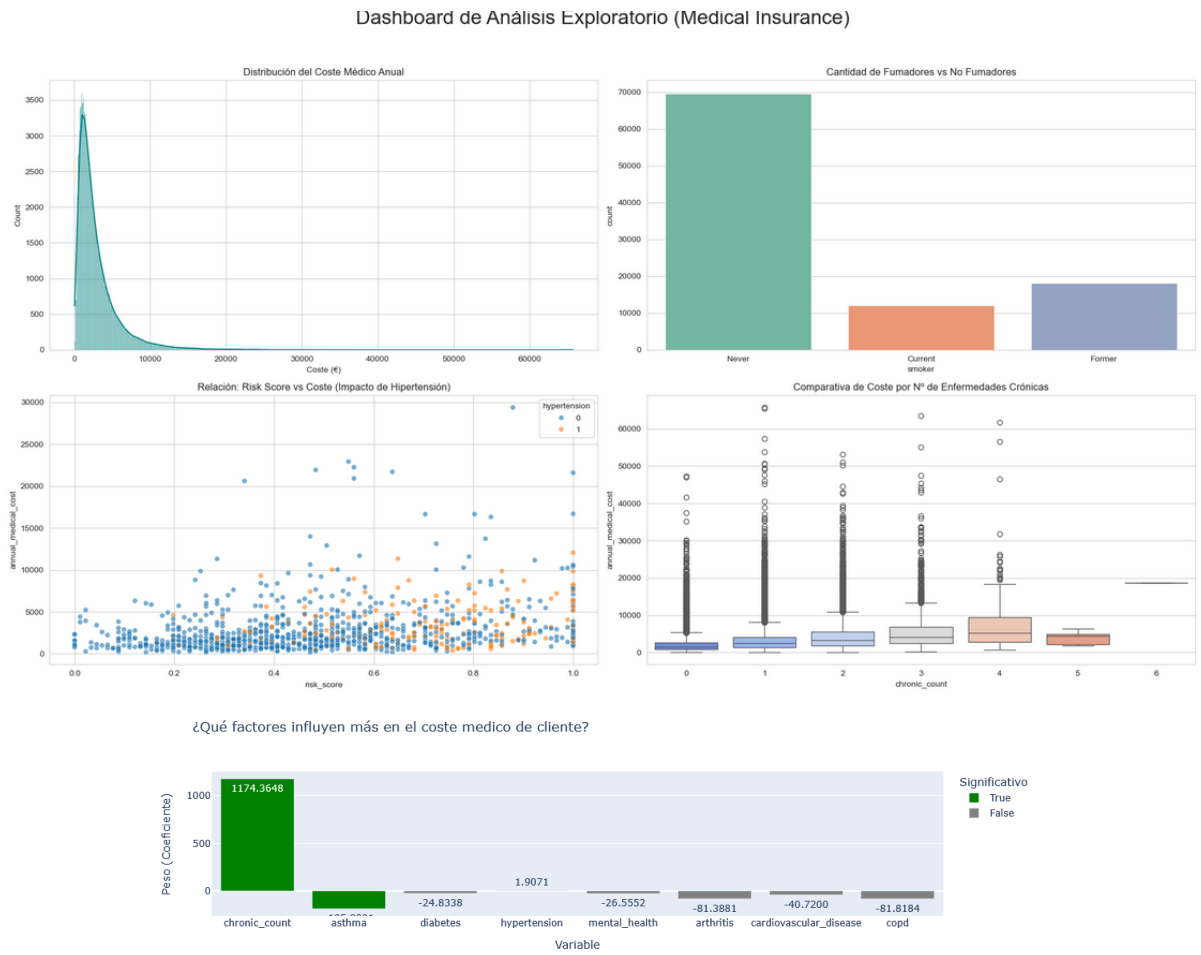
Nombre y Apellidos: PATRICIA SANCHEZ MARTIN
Github con notebook: https://github.com/Patricia-Sanchez-M/EXAMEN-FINAL-25-26

Nota: Por favor, seguir esta estructura para el documento

1. Resumen Ejecutivo

Escribe aquí tu resumen ejecutivo con el dashboard, hallazgos del análisis descriptivo que se observan el dashboard, insights del modelo predictivo y recomendaciones accionables basadas en los datos.



En la primera gráfica (arriba a la izquierda) vemos que la distribución de costes está sesgada a la derecha. Es decir, hay un grupo de pacientes reducido que está generando un gran coste anual para la aseguradora. El objetivo de este análisis es proponer formas de identificar a estos individuos y subir sus primas en función de esas variables críticas.

Proporcionalmente, como podemos ver en el gráfico de barras del dashboard, tenemos bastantes fumadores o antiguos fumadores. ¿Es esta una métrica clave para predecir el coste anual medio de un cliente? Mi análisis me lleva a pensar que hay otras mucho más importantes y que están siendo ignoradas, porque hay un menor número de casos.

La variable que a priori parece tener la mayor influencia con el coste anual medico por paciente es la edad del mismo. Sin embargo, un análisis gracias a la clusterización de los pacientes por edad ha revelado que esta 'cola' en la distribución que he comentado en el párrafo anterior no es debido a ancianos (>65), sino más bien a un grupo reducido de gente de mediana edad, en torno a los 50 años, que la compañía no está detectando con precisión su score de riesgo:

```
--- PERFIL DE LOS GRUPOS DETECTADOS ---
   Cluster        age  risk_score  annual_medical_cost  Cantidad_Pacientes
0        0  48.459381    0.502784          2310.011343               40461
1        1  29.735951    0.244072          1964.166144               28347
2        2  51.548508    0.708198         12009.584831                6133
3        3  65.140788    0.813268          3118.513121               25059
```

Tras realizar un modelo de regresión lineal para estimar el annual_medical cost en función del resto de variables del dataset (con una R^2 de más de 0.9), se ha hallado lo siguiente:

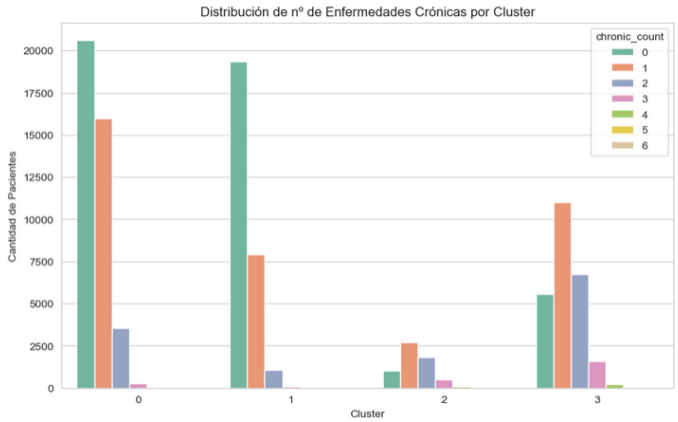
```
--- VARIABLES QUE MÁS SUBEN/BAJAN EL PRECIO ---
                  Variable   Coeficiente  Impacto_Absoluto
25           chronic_count  1.747163e+13      1.747163e+13
26            hypertension -8.727643e+12      8.727643e+12
35           mental_health -7.284317e+12      7.284317e+12
34               arthritis -6.742051e+12      6.742051e+12
27                diabetes -6.091385e+12      6.091385e+12
28                  asthma -5.118550e+12      5.118550e+12
30  cardiovascular_disease -4.781375e+12      4.781375e+12
29                    copd -4.038799e+12      4.038799e+12
3           household_size  3.897594e+12      3.897594e+12
4               dependents -3.448402e+12      3.448402e+12
```

Básicamente, que las variables que más afectan al coste anual de la aseguradora son principalmente patalogías actuales o históricas, como haber padecido cáncer en el pasado o tener hipertensión. Especialmente chronic_count.

Por otro lado, si realizamos igualmente un modelo de regresión lineal (R^2 = 0.97) para ver las variables que más lo están afectando, vemos lo siguiente, que la mayoría poco tienen que ver con las que vimos en el modelo anterior, son más sobre el estilo de vida de la persona, si está casada, consume alcohol…

```
--- VARIABLES QUE MÁS SUBEN/BAJAN EL PRECIO ---
                   Variable   Coeficiente  Impacto_Absoluto
3            household_size -1.926862e+09      1.926862e+09
61   marital_status_Married  1.838977e+09      1.838977e+09
4                dependents  1.704795e+09      1.704795e+09
71       alcohol_freq_Daily -1.209681e+09      1.209681e+09
73      alcohol_freq_Weekly -1.209681e+09      1.209681e+09
72  alcohol_freq_Occasional -1.209681e+09      1.209681e+09
74     alcohol_freq_missing -1.209681e+09      1.209681e+09
50              region_West  1.189715e+09      1.189715e+09
46           region_Central  1.189715e+09      1.189715e+09
49             region_South  1.189715e+09      1.189715e+09
```

Por ejemplo, si nos metemos un poco más en profundidad en los clusters de antes, vemos que el número de enfermedades crónicas mayores a cero es súper común en el cluster 2, que es en el que habíamos identificado que costaban mucho a la empresa todos los años. Esto pasa igual si vemos la proporción de pacientes que tienen diabetes, han tenido cáncer… en cada uno de los clusters. De hecho en el boxplot del dasboard podemos ver como aumenta el coste medio cuantas más enfermedades crónicas tenga el cliente.



Yo aquí entonces veo un problema claro por parte de la política de valoración de primas por parte de la compañía aseguradora que está siendo estudiada (asumo que la prima para el cliente se calcula usando el risk_score):

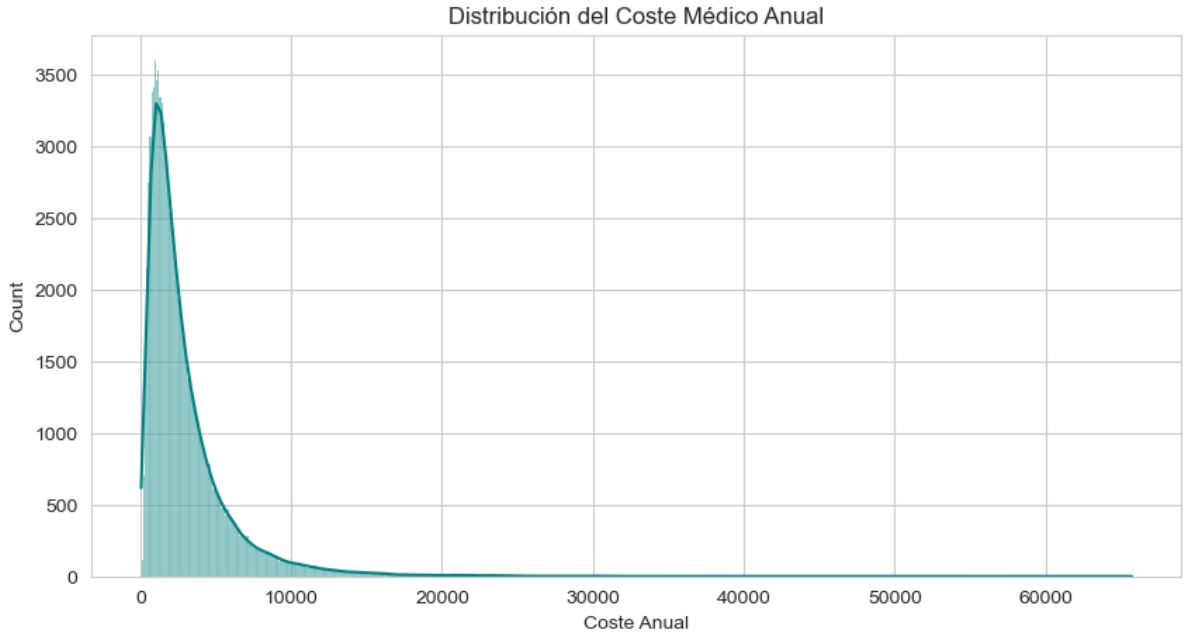
El cálculo de la prima se está realizando mayoritariamente basado en factores de estilo de vida del cliente PERO el coste anual médico que genera esa persona no depende tanto de su estilo de vida sino de si tiene o ha tenido alguna patología médica en el pasado. Esto lo vemos en el scatter plot del dasboard, actualmente, no hay una correlación fuerte entre risk_score y annual_medical_cost, y, como parámetro clave que se usa para calcular la prima, considero que sí que deberían estar mucho más correlacionadas.

Por tanto, recomiendo ajustar el risk_score para dar más peso a estos elementos de historial médico y elevar la prima si se presentan en un cliente.

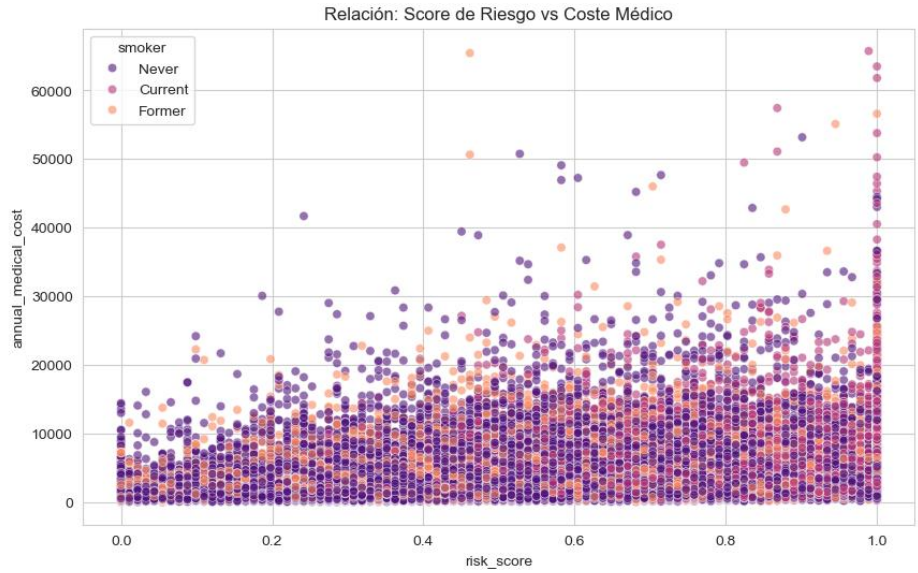
Además, dado que se ha identificado al cluster que está generando mayor coste anual médico y que es de mediana edad, poner más énfasis en investigar historiales médicos de personas de esa edad que quieran contratar un seguro, para poder cobrarles primas más altas y compensar los costes médicos que le cuestan a la compañía aseguradora cada año.

2. Gráficas del análisis exploratorio y breve explicación de cada una

Para situarme en contexto, he elaborado un histograma sencillo con la distribución del coste médico anual, que es la variable que voy a investigar. Como se puede ver, la mayoría de los asegurados (alrededor de 3500) tienen un coste anual de alrededor de 3000.

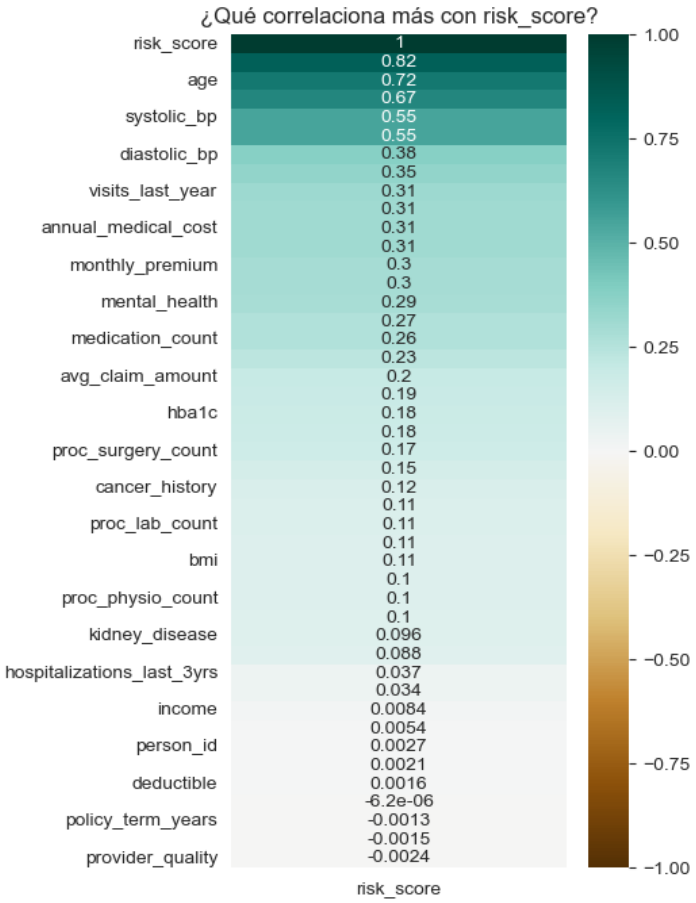


Siguiente, he querido ver cómo afecta el consumo de tabaco al 'risk_score' que te asigna la compañía aseguradora, y finalmente cómo se relaciona también con el 'annual_medical_cost':

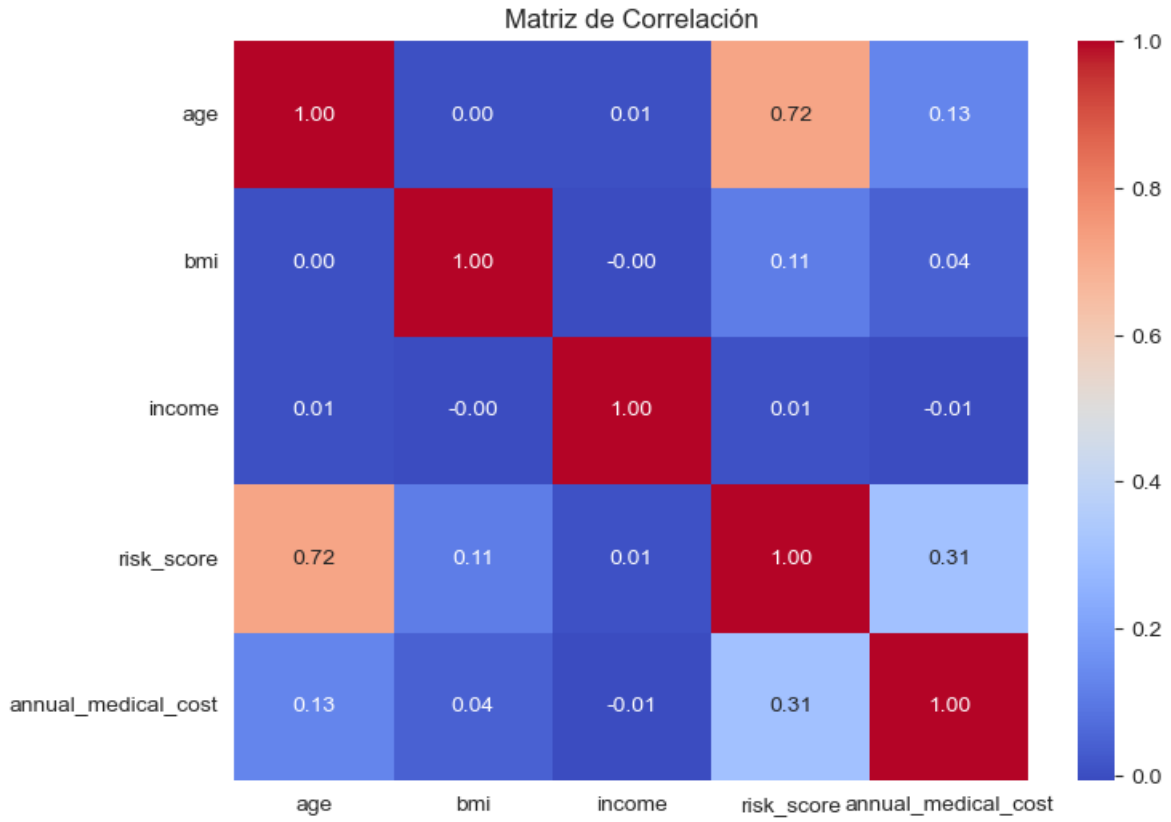


Aquí podemos ver como que los outliers más significativos (arriba a la derecha) que tienen tanto un alto risk_score como alto annual_medical_cost, son mayoritariamente fumadores en la actualidad.

Siguiendo por esta línea de qué factores afectan al risk_score de un paciente, he hecho este gráfico de correlación, que nos da las variables más correladas con el mismo. Tal y como sospecha la compañía, depende mucho de la edad.



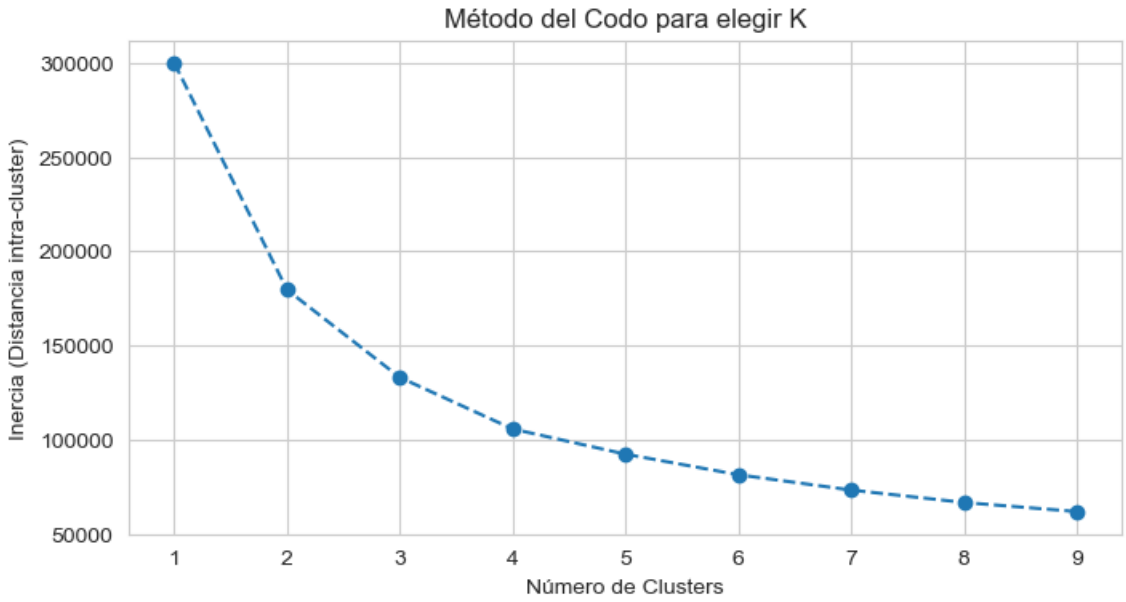
Luego he hecho una matriz de correlación incluyendo el risk_score con varias variables de hábitos y estilo de vida.



Además, se ha realizado una clusterización k-means con:

- Annual_medical_cost: que estamos analizando
- Risk_score: otra variable que queremos entender de qué depende
- Edad: hemos visto por la correlacion que afecta mucho al risk_score

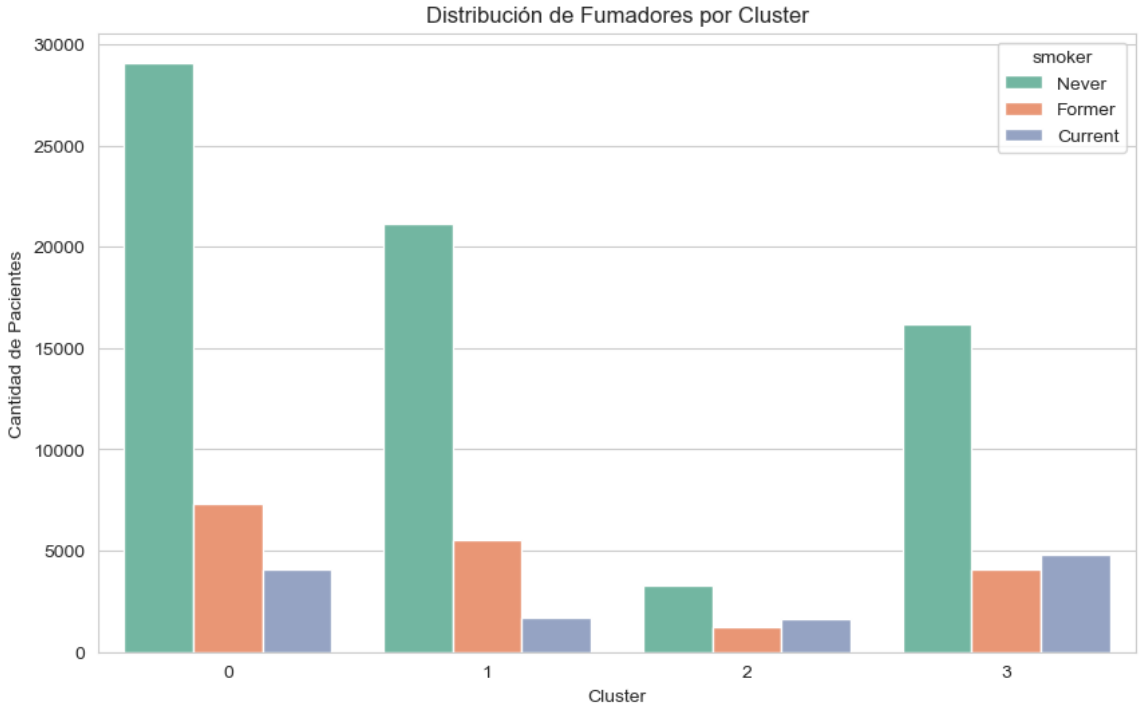
En esta clusterización se ha determinado con el método del codo que el número óptimo de clusters para dividir a los pacientes es 4.




```
--- PERFIL DE LOS GRUPOS DETECTADOS ---
   Cluster        age  risk_score  annual_medical_cost  Cantidad_Pacientes
0        0  48.459381    0.502784          2310.011343               40461
1        1  29.735951    0.244072          1964.166144               28347
2        2  51.548508    0.708198         12009.584831                6133
3        3  65.140788    0.813268          3118.513121               25059
```

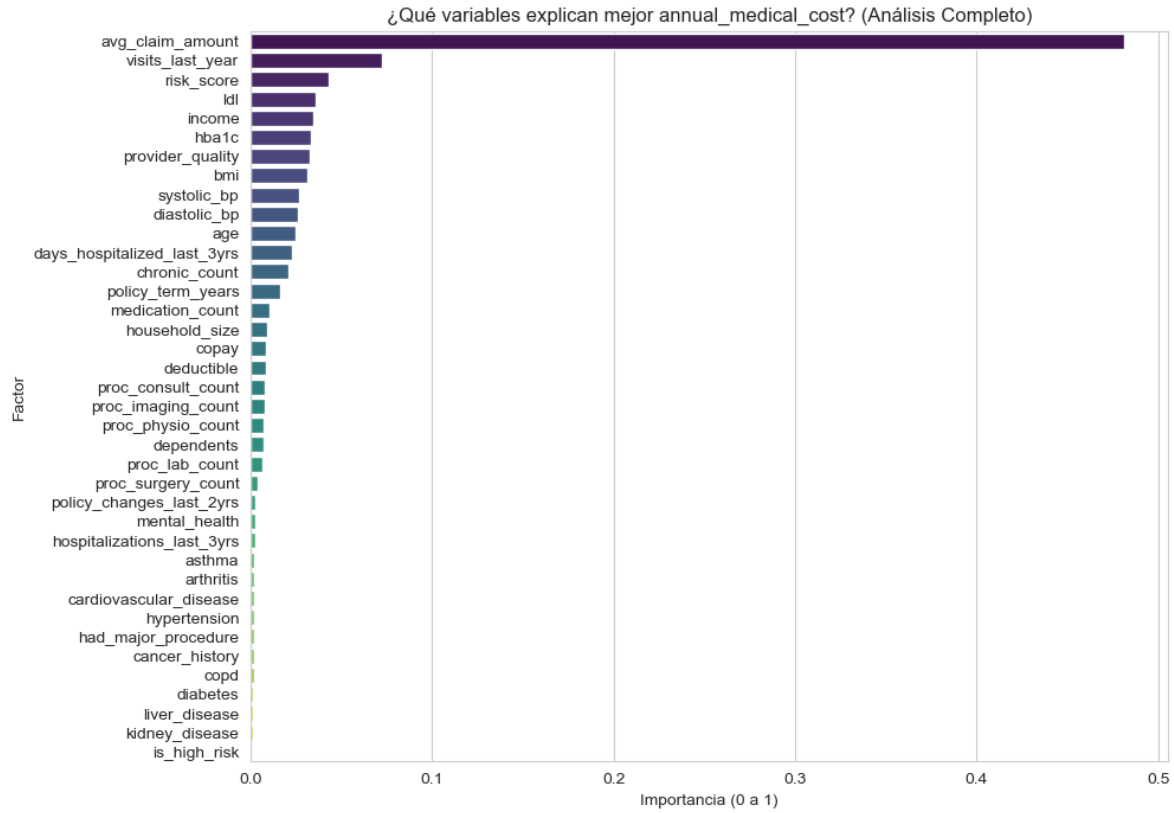
Se han creado 4 grupos como veis arriba. Lo que me llama la atención es el cluster 2, es el más pequeño además. Es un grupo de gente de mediana edad, con un risk_score muy alto (pero NO debido a la edad, tengo que investigar a qué) y con el mayor annual_medical_cost!! Esto es una sorpresa, ya que por todo el analisis anterior pensaria que el grupo con mayor coste médico sería gente de edad avanzada.

Si diseccionamos cada cluster para ver cuántos fuman en cada cluster, tenemos esto:



Esto me lleva a pensar que fumares algo bastante relevante, porque la proporcion de fumadores en el cluster 2 respecto a los otros es más significativa.

He decidido usar un random forest para que me de los pesos de cada variable en el annual_medical_cost, quitando las cosas obvias que sea como la prima pagada, la cantidad de emergencias que ha tenido…. y me ha salido lo siguiente:



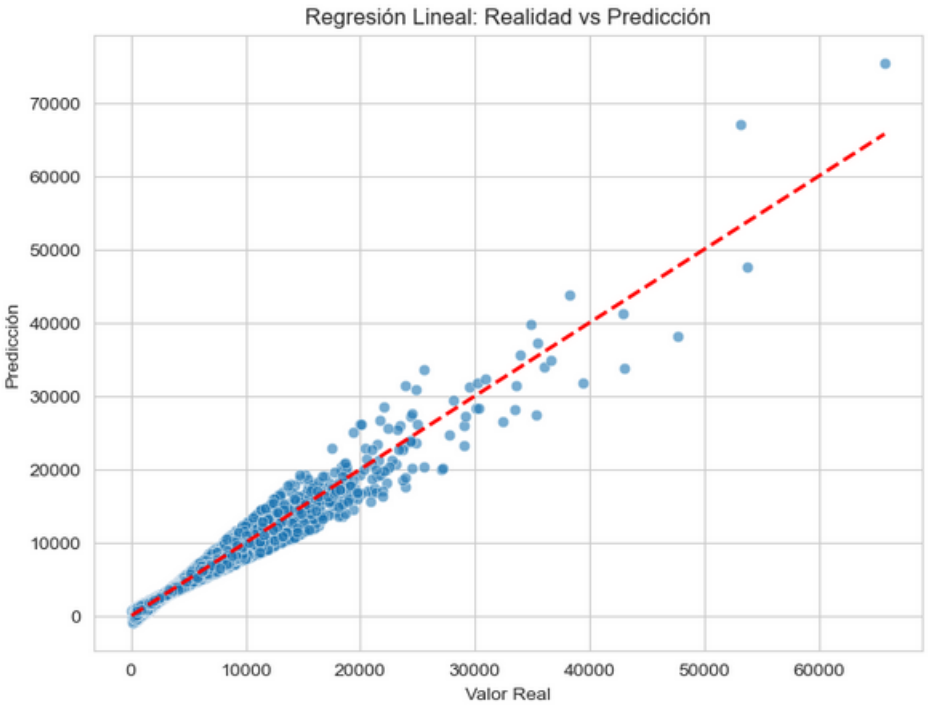
3. Modelo predictivo explicado y con tablas

Este es un modelo de regresión lineal con todas las variables numéricas.

Tiene una R^2 muy alta, por lo que es bastante bueno.

En él se determina que las variables que más suben/bajan el precio. Así que ahora voy a explorar por ahí.

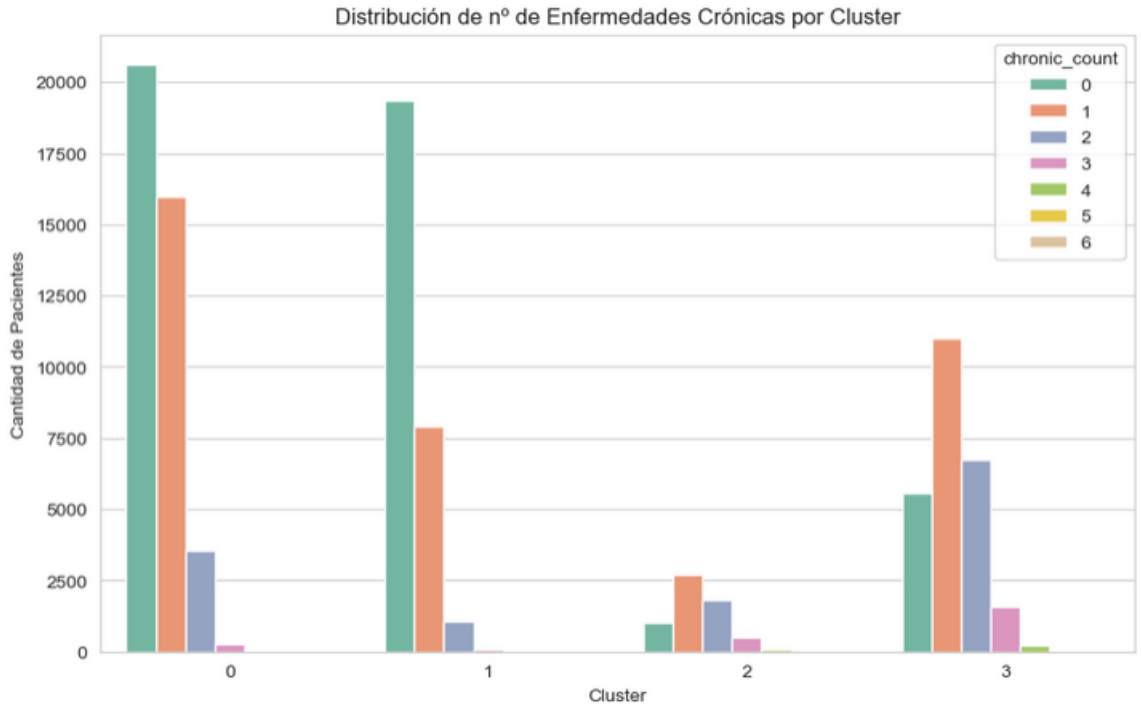
```
Entrenando Regresión Lineal...

--- RESULTADOS REGRESIÓN LINEAL ---
R2 Score (Calidad del modelo 0-1): 0.9664
MAE (Error medio absoluto): 319.92 euros
RMSE (Raíz del error cuadrático medio): 574.87 euros
```




```
--- VARIABLES QUE MÁS SUBEN/BAJAN EL PRECIO ---
                 Variable   Coeficiente  Impacto_Absoluto
25          chronic_count  1.747163e+13      1.747163e+13
26           hypertension -8.727643e+12      8.727643e+12
35          mental_health -7.284317e+12      7.284317e+12
34              arthritis -6.742051e+12      6.742051e+12
27               diabetes -6.091385e+12      6.091385e+12
28                 asthma -5.118550e+12      5.118550e+12
30  cardiovascular_disease -4.781375e+12      4.781375e+12
29                   copd -4.038799e+12      4.038799e+12
3          household_size  3.897594e+12      3.897594e+12
4              dependents -3.448402e+12      3.448402e+12
```

Si ahora hacemos el análisis por cluster en función del nº de enfermedades crónicas, que es la variable que ha salido más importante en el modelo de regresión lineal, pues vemos que en los dos primero clusters, abunda la gente ‘sana’, mientras que, en los otros dos, hay bastante más gente con enfermedades cronicas frente a la gente sana.

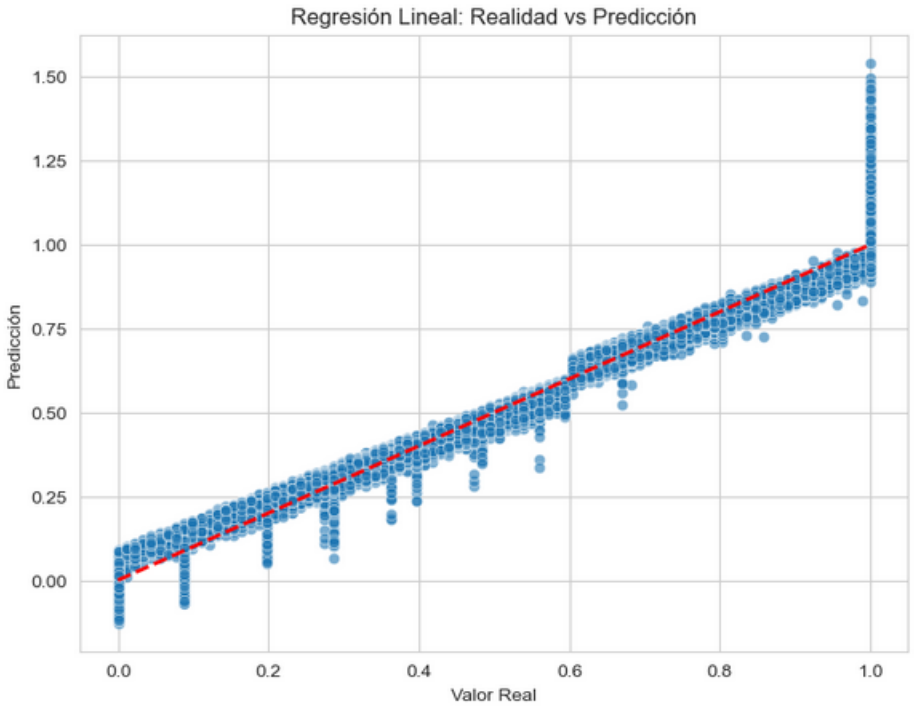
Lo que pasa es que, de nuevo, vemos lo inusual en el cluster 2 porque es gente ‘relativamente joven’ de media 50 años que tiene estas patologías, frente al cluster 3 que tiene metidos a los clientes más mayores, con lo cual es más normal encontrarnos a un abuelo con mayor numero de enfermedades crónicas.



Sin embargo, volviendo a la figura de antes para ver de qué depende el risk_score que la aseguradora asigna al paciente, este parámetro no salía como importante, por lo que recomiendo darle más peso.

De hecho, si hacemos el mismo modelo de regresión lineal para ver qué variables afectan más al risk_score, vemos lo siguiente:

```
Entrenando Regresión Lineal...

--- RESULTADOS REGRESIÓN LINEAL ---
R2 Score (Calidad del modelo 0-1): 0.9715
MAE (Error medio absoluto): 0.03 euros
RMSE (Raíz del error cuadrático medio): 0.04 euros
```




```
--- VARIABLES QUE MÁS SUBEN/BAJAN EL PRECIO ---
                  Variable   Coeficiente  Impacto_Absoluto
3           household_size -1.926862e+09      1.926862e+09
61   marital_status_Married  1.838977e+09      1.838977e+09
4               dependents  1.704795e+09      1.704795e+09
71       alcohol_freq_Daily -1.209681e+09      1.209681e+09
73      alcohol_freq_Weekly -1.209681e+09      1.209681e+09
72  alcohol_freq_Occasional -1.209681e+09      1.209681e+09
74     alcohol_freq_missing -1.209681e+09      1.209681e+09
50             region_West  1.189715e+09      1.189715e+09
46          region_Central  1.189715e+09      1.189715e+09
49            region_South  1.189715e+09      1.189715e+09
```

Lo que veo es que está muy basado en hábitos y estilo de vida. De nuevo llego a la conclusión de que el risk_score está calculado basado en parámetros que no tienen tanto que ver con luego el verdadero coste anual que ese paciente genera.

Yendo un poco más allá en esta línea de que no es tanto los hábitos y estilo de vida (deporte, fumadores, BMI….) sino historial médico.

Si vemos el historial médico de uno a uno de los clusters….

```
--- ANÁLISIS DEL CLUSTER 0 (Total: 40461 pacientes) ---
Coste medio: 2310.01 €
Edad media: 48.5 años

Conteo de enfermedades en este grupo:
hypertension      6983
diabetes          2823
asthma            1920
cancer_history     683
```

```
--- ANÁLISIS DEL CLUSTER 1 (Total: 28347 pacientes) ---
Coste medio: 1964.17 €
Edad media: 29.7 años

Conteo de enfermedades en este grupo:
hypertension      2986
diabetes          1152
asthma             824
cancer_history     288


 --- ANÁLISIS DEL CLUSTER 2 (Total: 6133 pacientes) ---
 Coste medio: 12009.58 €
 Edad media: 51.5 años

 Conteo de enfermedades en este grupo:
 hypertension      2225
 diabetes           989
 asthma             640
 cancer_history     266


 --- ANÁLISIS DEL CLUSTER 3 (Total: 25059 pacientes) ---
 Coste medio: 3118.51 €
 Edad media: 65.1 años

 Conteo de enfermedades en este grupo:
 hypertension      8151
 diabetes          3629
 asthma            2503
 cancer_history     914
```

El 2 es, en proporción al menor número de personas, el grupo que genera más gasto para la compañía y el que concentra a los individuos con más enfermedades y patologías.

Si hacemos un modelo lineal SOLO con patologías médicas, chronic_count es la que más pesa con diferencia: