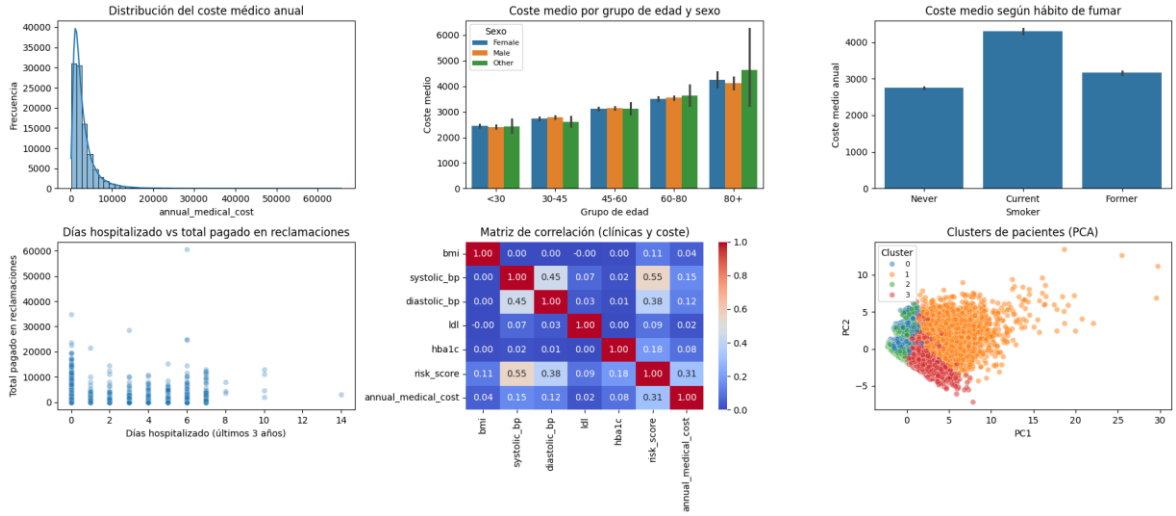
Nombre y Apellidos: Alonso Lidon Gonzalo
Github con notebook: https://github.com/GonAlonsoLid/runner-air-planner/blob/main/DAVD_Examen_final_2025_2026.ipynb

1. Resumen Ejecutivo

Máximo 2 páginas

Este análisis explora los principales factores clínicos, demográficos y de comportamiento que influyen en el coste médico anual y en el riesgo sanitario de una población de pacientes. A partir de un estudio exploratorio detallado, se observa que la mayoría de los pacientes presentan costes moderados, mientras que un pequeño grupo concentra gastos extremadamente altos, lo que genera una clara distribución asimétrica. Este patrón confirma la existencia de pacientes con necesidades complejas cuyo impacto económico es muy significativo.



El análisis por segmentos revela que el envejecimiento y el hábito de fumar son dos de los impulsores más consistentes del aumento del coste. Los grupos de mayor edad tienen incrementos en el gasto sanitario independientemente del sexo, mientras que los fumadores presentan los costes más elevados en comparación con exfumadores o quienes nunca han fumado.

La correlación entre variables clínicas muestra que no existe un único indicador fuerte aislado, aunque el risk_score destaca como el predictor clínico más alineado con el coste. Esto sugiere que las características de salud deben interpretarse en conjunto y no de manera individual.

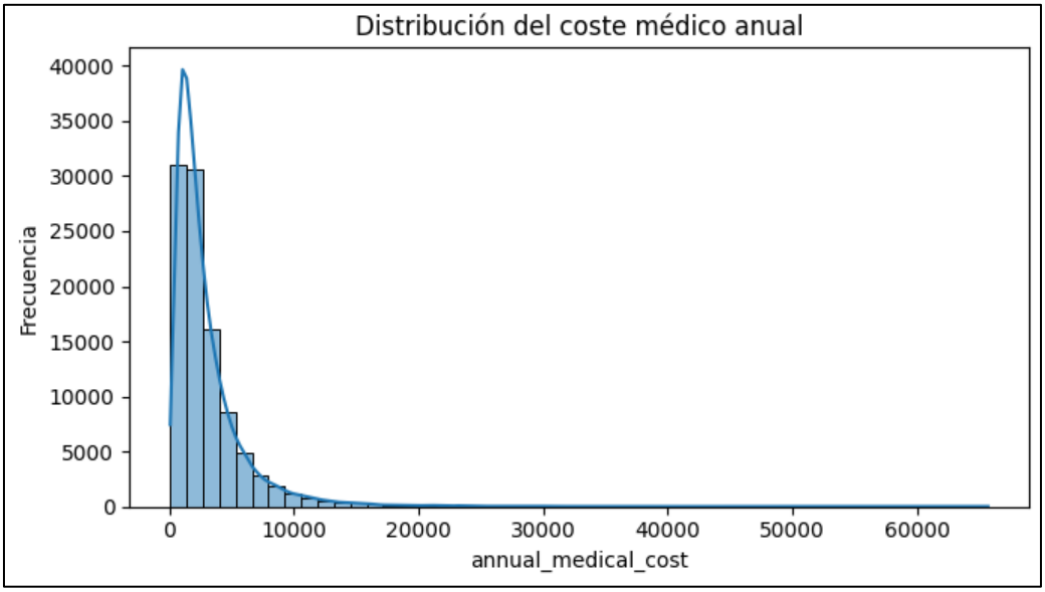
El clustering permitió identificar cuatro perfiles de pacientes diferenciados, desde pacientes jóvenes y clínicamente estables, hasta grupos con alta carga de enfermedades crónicas, muchas hospitalizaciones y costes muy elevados. La visualización mediante PCA confirma una separación clara entre estos segmentos, confirmando la diferencia entre grupos.

Finalmente, se desarrolló un modelo predictivo de regresión logística para clasificar pacientes en riesgo alto o bajo. El modelo alcanza una precisión muy alta y confirma que las variables que más influyen en el riesgo son: ser fumador actual, la edad y la carga crónica.

Todo lo anterior potencia el fomentar los hábitos saludables entre la población para reducir las clases de riesgo y así el coste medio anual por paciente, con el fin de tener una sociedad mas sana y con una esperanza de vida todavía mas alta de la actual.

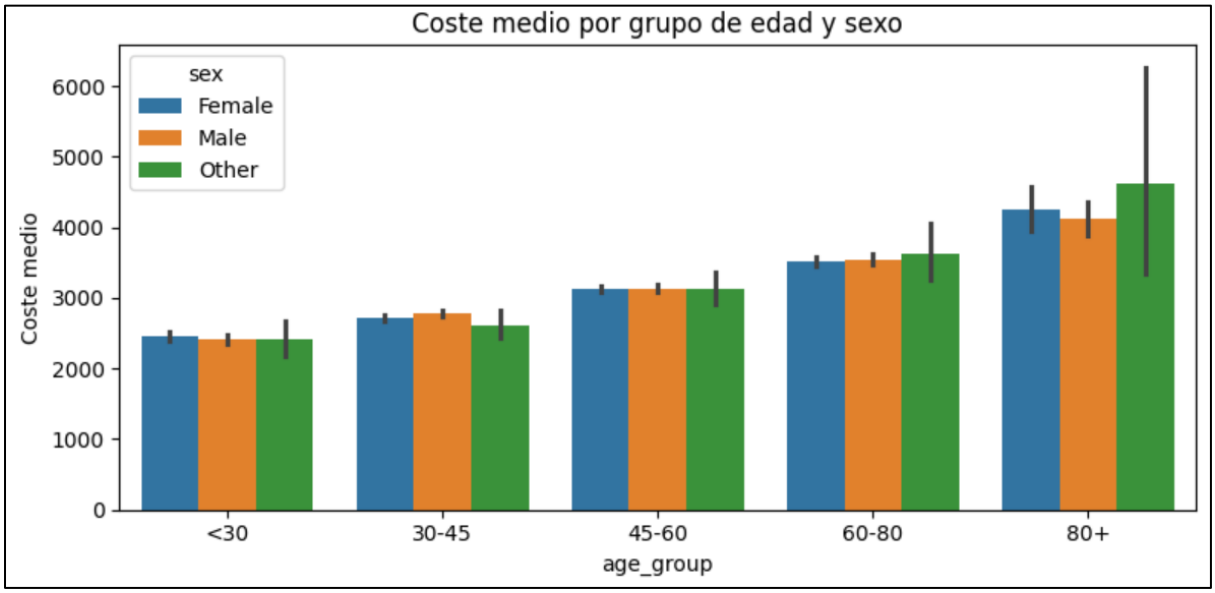
2. Gráficas del análisis exploratorio y breve explicación de cada una

Distribución del coste medico anual



El gráfico muestra que la mayoría de las personas tienen gastos médicos anuales bastante bajos, mientras que solo unos pocos llegan a tener costes muy altos, lo que genera esa cola larga hacia la derecha. Algo a tener en cuenta es que estos pocos casos extremos, aunque son raros, tienen un impacto enorme en el gasto total, por lo que podrían ser clave para diseñar estrategias de prevención o planes de seguros más eficientes.

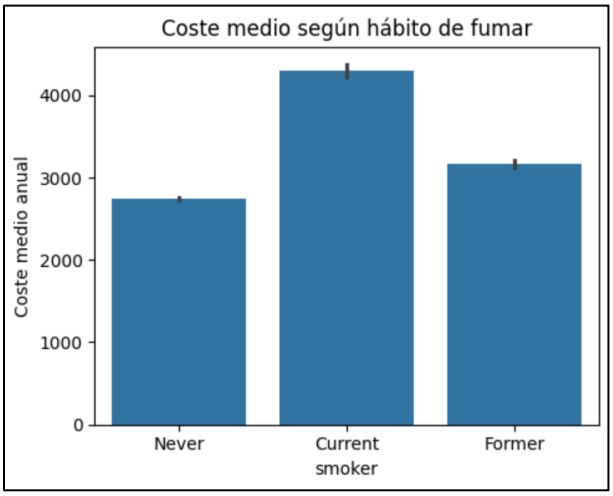
Coste medio por grupo de edad y sexo



Con esta gráfica podemos ver claramente la distribución del coste medio por sexo y grupos de edad. Gracias a estas barras somos capaces de entender cómo se reparte el gasto médico entre distintos segmentos demográficos. A simple vista se nota que, cuanto más avanza la edad, más aumentan los gastos médicos sin importar el sexo.

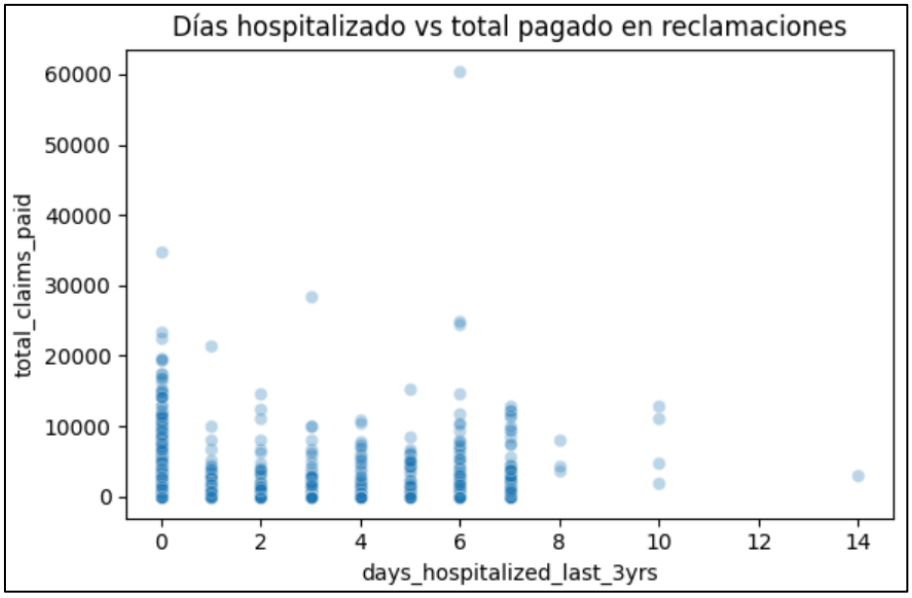
Las diferencias entre hombres, mujeres y “other” son pequeñas en los grupos jóvenes, pero empiezan a hacerse más visibles a partir de los 60 y se notan especialmente en el grupo de 80+, donde los costes se disparan. Un aspecto para tener en cuenta es que el aumento del gasto con la edad es tan consistente que demuestra que la variable más determinante del coste médico parece ser el envejecimiento. Esto es importante a la hora de considerar el desarrollo del modelo de *machine learning* más adelante.

Coste medio de fumadores



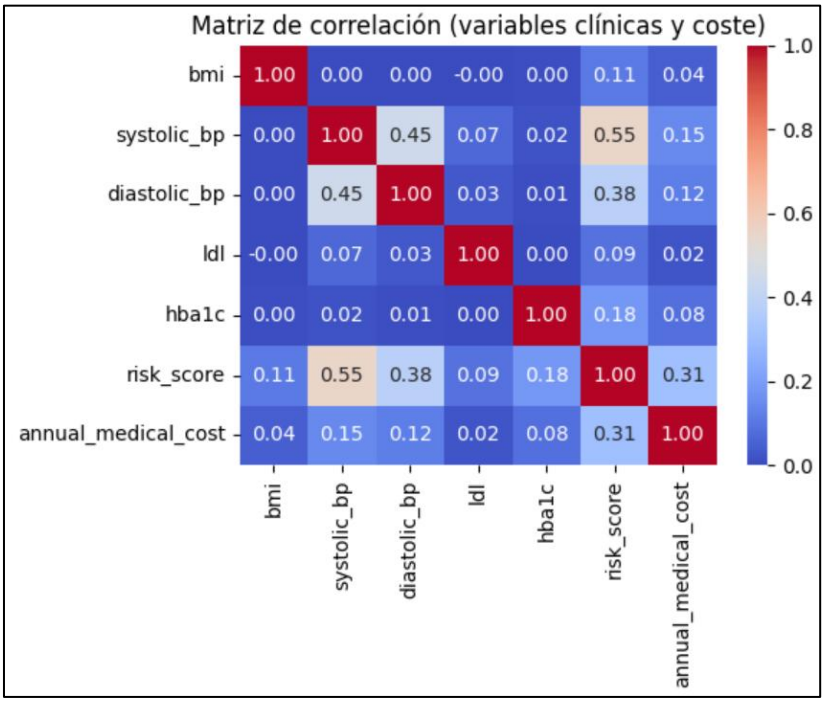
A la hora de determinar el coste médico es importante analizar el impacto del hábito de fumar, ya que esta variable está presente en nuestro dataset. Como podemos observar en la gráfica, las personas que fuman actualmente (*current*)son las que presentan el coste médico medio más alto, superando con bastante diferencia a los que nunca han fumado o a los que antes fumaban. Por otro lado, quienes nunca han fumado tienen el gasto más bajo, mientras que los exfumadores quedan en un punto intermedio.

Días hospitalizado vs total pagado en reclamaciones



Viendo esta nube de puntos, parece que detrás del coste total hay más variables en juego que solo el tiempo en el hospital, ya que el numero de reclamaciones puestas no crece con el número de días ingresado.

Matriz de correlación de las variables clínicas + coste



Por último, si analizamos solo las variables clínicas de nuestro dataset contra el coste medio anual, podemos observar que obtenemos una mayor relación entre el coste medio y el score de riesgo. De nuevo, esto apunta a que aquellos pacientes que tienen enfermedades o malos hábitos de salud tienden a generar gastos médicos más elevados, ya que requieren más seguimiento, y más tratamientos.

Clustering

Para la parte de clustering, en primer lugar, tenemos que seleccionar las variables con las que queremos realizar la segmentación de los datos. En este caso hemos escogido:

- Edad
- BMI
- Chronic count
- Visit last year
- Hospitalizations last three years
- Annual Medical Cost
- Total claims paid

Este conjunto de variables, por lo visto anteriormente, se han cogido por su carácter demográfico, así como la explicabilidad de estas frente al coste medio anual visto en las visualizaciones anteriores.

Usando un algoritmo de K-means obtenemos lo siguiente:

cluster	age	bmi	chronic_count	visits_last_year	hospitalizations_last_3yrs	annual_medical_cost	total_claims_paid
0	47.97	31.17	0.37	1.22	0.07	2355.14	645.73
1	51.95	27.35	1.48	3.46	0.35	12266.22	8302.30
2	42.64	22.93	0.37	1.27	0.07	1988.60	574.80
3	52.99	26.89	1.57	3.55	0.11	3513.41	2149.54

1. Perfil 0: Riesgo moderado y con un BMI relativamente alto

Este grupo reúne a pacientes de mediana edad con un BMI alto, lo que indica sobrepeso u obesidad. Aunque no tienen muchas enfermedades crónicas y apenas usan el sistema sanitario, este exceso de peso los sitúa en un riesgo futuro. Su coste actual es moderado, pero podrían evolucionar hacia perfiles más costosos.

2. Perfil 1: Sobrepeso, coste y enfermedades crónicas

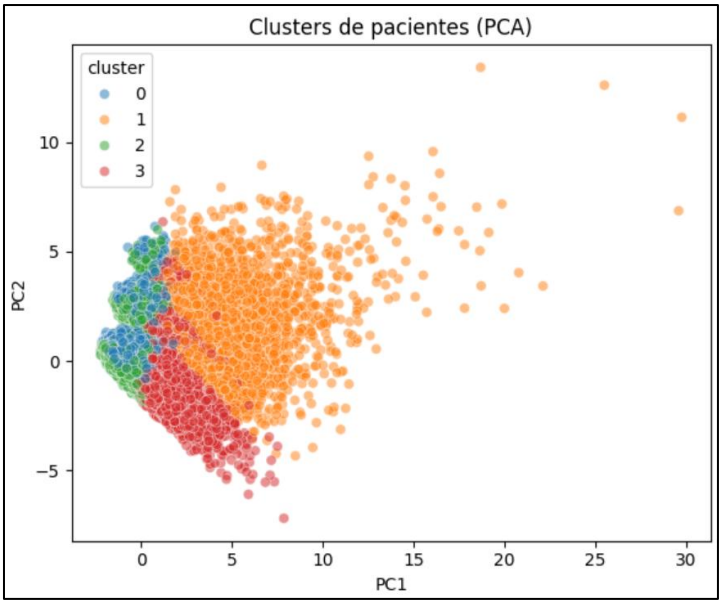
Aquí encontramos a los pacientes más costosos. Presentan varias enfermedades crónicas, más visitas médicas y más hospitalizaciones que ningún otro grupo. Esto se traduce en un coste anual muy elevado.

3. Perfil 2: Jóvenes

Este es el grupo más sano: pacientes relativamente jóvenes, con un BMI normal y casi sin enfermedades crónicas. Hacen poco uso del sistema sanitario y generan el coste más bajo de todos los clústeres. La clave aquí sería identificar las características que definen este perfil y promover hábitos saludables y prevención.

4. Perfil 3: Alto uso del sistema

Los pacientes de este clúster se parecen a los del 0 en cuanto a BMI, pero presentan más visitas y más hospitalizaciones, lo que eleva su coste anual. No llegan al nivel crítico del clúster 1, pero sí muestran un uso mayor del sistema que justifica que sean vigilados más de cerca.



Como se puede ver en la gráfica PCA existe una clara distinción entre los diferentes perfiles, con una similitud muy cercana en el caso de los clústeres 0 y 2 como se ha comentado anteriormente por las métricas de salida. Se trata de clústeres con diferencias pequeñas, pero notables a efectos de sus gastos médicos anuales.

3. Modelo predictivo explicado y con tablas

Para predecir si un paciente es de alto riesgo construimos un modelo de regresión logística que combina tanto variables numéricas como categóricas. Estas son las siguientes:

"age", "sex", "bmi", "smoker",
"chronic_count",
"visits_last_year", "hospitalizations_last_3yrs",
"plan_type", "network_tier"

Primero preparamos los datos con un *preprocesador* que estandariza las variables numéricas y convierte las categóricas en variables dummies mediante *one-hot encoding*.

Después entrenamos el modelo con una división entre entrenamiento (80%) y test (20%) para mantener el equilibrio entre clases.

```
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=0,
)
```

Una vez entrenado, ejecutamos la predicción sobre los datos de test y obtenemos la siguiente matriz de confusión.

```
              precision    recall  f1-score   support

           0       0.98      0.98      0.98     12754
           1       0.96      0.96      0.96      7246

    accuracy                           0.97     20000
   macro avg       0.97      0.97      0.97     20000
weighted avg       0.97      0.97      0.97     20000

[[12453   301]
 [  309  6937]]
```

Los resultados del modelo son muy buenos, con un 97% de precisión de media y una *accuracy* del 97%.

La matriz de confusión nos confirma que el modelo esta acertando en la mayoría de los casos (12453+6937). Esto sugiere que el modelo está aprendiendo patrones muy fuertes en los datos, aspecto que se reafirma en los coeficientes de cada variable.

	feature	coef
7	smoker_Current	11.266164
0	age	10.919302
2	chronic_count	9.923237
8	smoker_Former	-5.909854
9	smoker_Never	-5.840364
1	bmi	1.926929
4	sex_Female	-0.211054
12	plan_type_POS	-0.188032
6	sex_Other	-0.154657
13	plan_type_PPO	-0.142147
5	sex_Male	-0.118343
11	plan_type_HMO	-0.101689
10	plan_type_EPO	-0.052186
3	visits_last_year	0.000181

Como podemos ver, el modelo deja bastante claro que el hábito de fumar es el factor que más empuja el riesgo de un paciente. Ser fumador aparece con un coeficiente muy alto, indicando que este grupo tiene mucha más probabilidad de ser clasificado como “alto riesgo”. Esto encaja perfectamente con lo que se ve en el análisis exploratorio: el tabaco suele estar detrás de una larga lista de complicaciones de salud, por lo que su peso en el modelo tiene todo el sentido. Por otro lado, los coeficientes negativos asociados a “exfumadores” o a quienes nunca han fumado muestran cómo adoptar hábitos saludables reduce de forma notable ese riesgo.

También destaca la edad, que aparece como el segundo factor más importante. A medida que una persona envejece, es más probable que acumule enfermedades crónicas o necesite más atención sanitaria, por lo que su contribución al riesgo es normal. En la misma línea, el *chronic_count* confirma que la carga de enfermedad es un pilar clave en la

predicción: cuantos más problemas crónicos tenga un paciente, mayor es la probabilidad de acabar en el grupo de alto riesgo.

El resto de variables, como el BMI, el tipo de plan o el sexo, tienen un impacto mucho más pequeño. Esto nos indica que, aunque puedan influir algo, no son los elementos que realmente marcan la diferencia.