Nombre y Apellidos: Marian Moreno Nieto
Github con notebook:
Nota: Por favor, seguir esta estructura para el documento

1. Resumen Ejecutivo

Máximo 2 páginas

En este **dashboard** tengo 4 graficas.

Grafica 1: Permite ver la cantidad de riesgo del total de mis pacientes separando por genero masculino y femenino, aquello le permitirá saber al analista de datos el total de riesgo de su portfolio para hacerse una idea de que magnitudes se esta hablando. Se separa por genero masculino y femenino ya que es la primera componente genética y la que afectara a la hora de comunicaciones y también a la hora de que enfermedades sufres, donde ocurren los gastos.

Grafica 2: Permite ver la importancia de los coeficientes según el modelo a la hora de analizar los gastos totales de cada paciente; se actualiza según el modelo usado con coef_table para que en el caso de encontrar un modelo mejor al hecho en el ejercicio de mas información. Ahora mismo el valor principal es el risk scroe ya que a mayor riesgo, el paciente seguramente incurra en mas gastos.

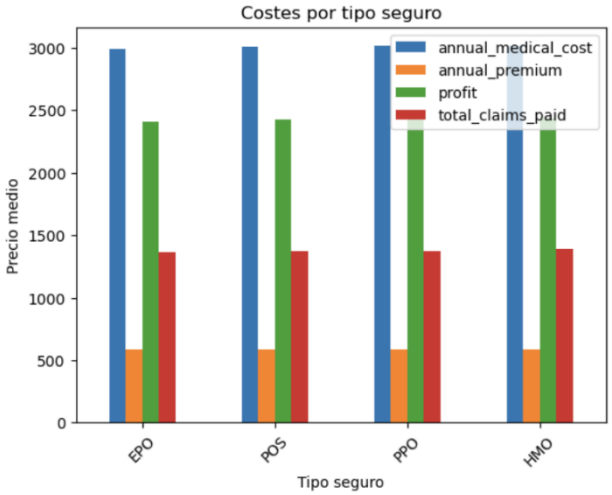
Gráfica 3: Se analiza los dias que se ha estado en el hospital según su edad (y separando por colores la cantidad de alcohol ingerida). Además para quitar posibles “bias” hemos puesto el filtro donde 0 indica que no hay tenido un procedimiento medico gradne y 1 que si; de esta manera no estña afectando al numero de dias en el hospital si ha tenido un procedimiento importante sino que los analizamos como gráficas separadas. Esto permite al analista observar como cuando se ha ingerido mas alchol; han sufrido una operación grande.

Gráfica 4: Se filtran por 0 (low risk) y 1 (high risk) ya que veíamos en los coeficientes y en las correlaciones que el risk es el factor que afecta principalmente al precio; por lo que a la hora de evaluar debemos ver low y high risk por separado. Además estamos viendo el risk score según si es fumador o no; entonces aquellos que no son fumadores tiene un risk score mucho mas pequeño (e incluso los que han dejado de fumar también)

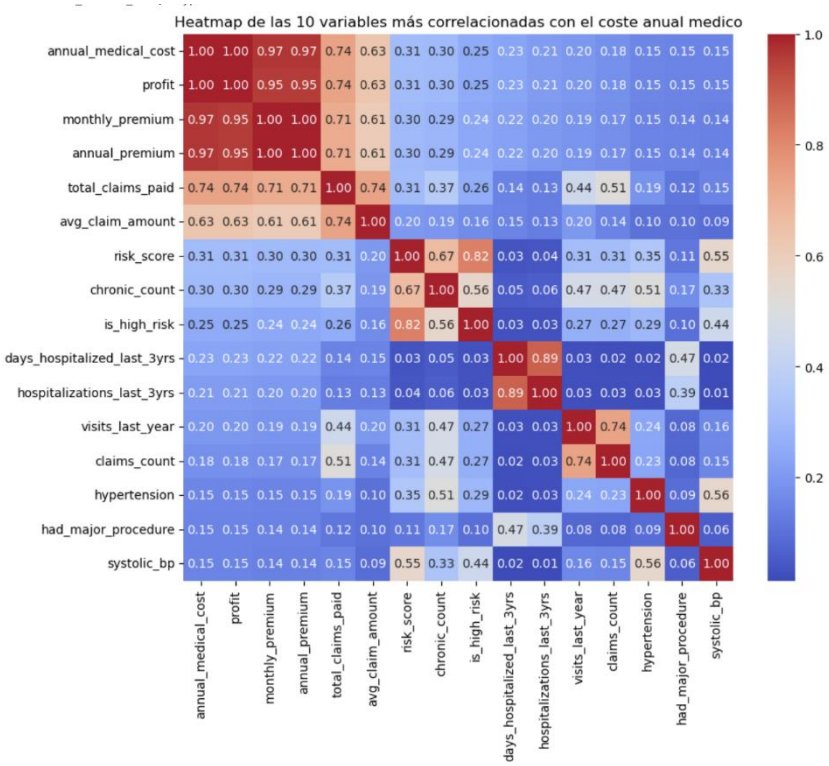
En conclusión: La variable que mas afecta al coste es el risk + dias en el hospital y estas a su lado se ve una clara diferencia al separar por alcohol / tabaco como afecta. Este dshboard permite identificar patrones y aquellos pacientes que incurren en una de las dos, separando en cada uno de los dos casos con filtros las dos variables que mas afectan a cad uno (operación grande o high/low risk) para quitar bias.

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

Copia aquí tus gráficas y explícalas. Mínimo 6.



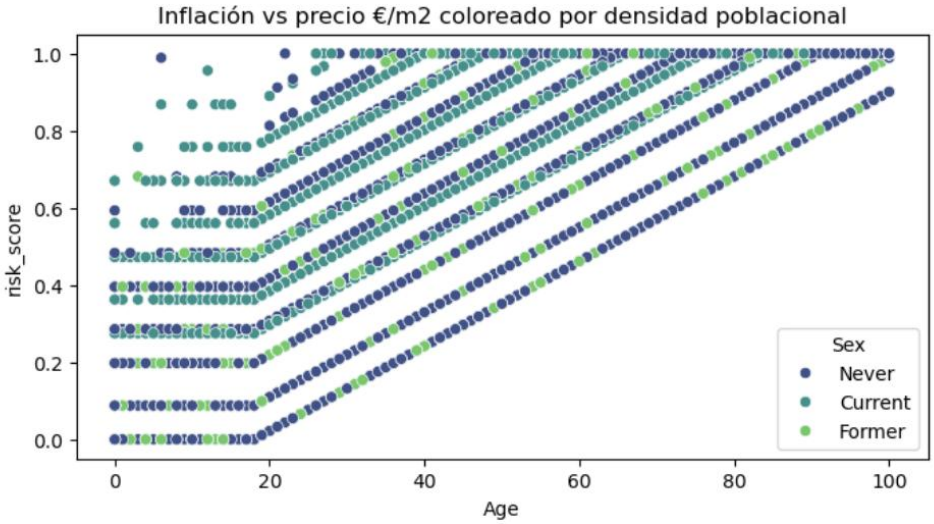
Podemos ver como el tipo de seguro no afecta ya que la media de costes del paciente; , la media pagada a modo de prima a la aseguradora y el total de claims pagadas es mnuy parecida en todos los tipos de seguro (Además el beneficio (coste anual-prima anual) podemos)



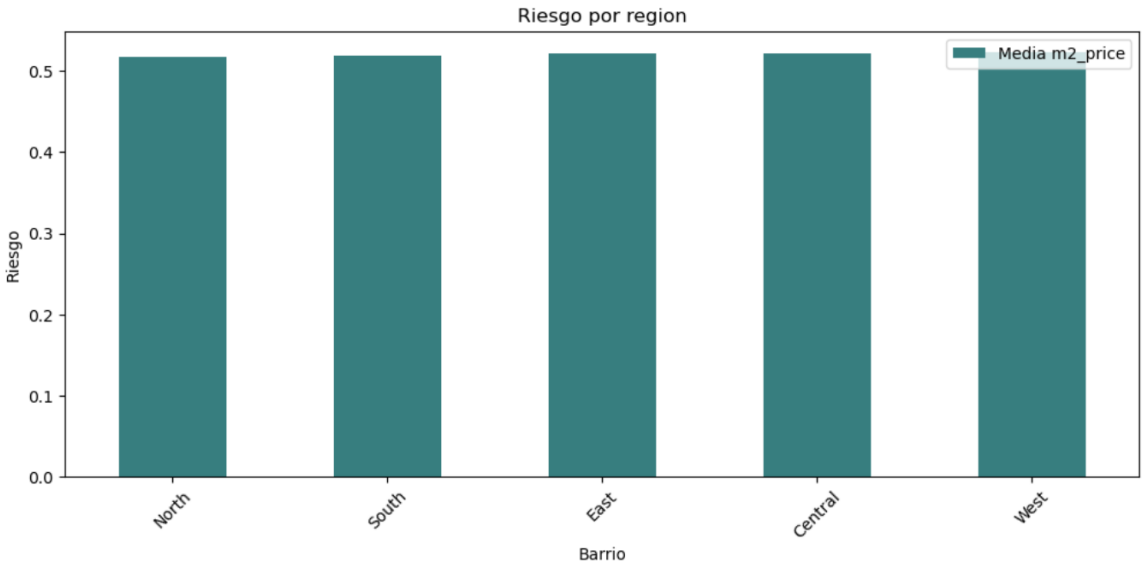
Quitando las variables obvias con correlación (todas las relacionadas con costes y pagos encontramos que las variables que mas van a afectar al coste (es decir correlacionadas positivamente con el coste anual por paciente son:

Por lo tanto vemos que hay enfermedades que llevan a un mayor gasto como puede ser arthirits, hipertension, dyastolis y systemic bp frente a otras que no tanto y que tiene mucho mas peso por ejemplo el avg pagado por factura (avg_claim_amount a que hayas

tenido muchas claims)

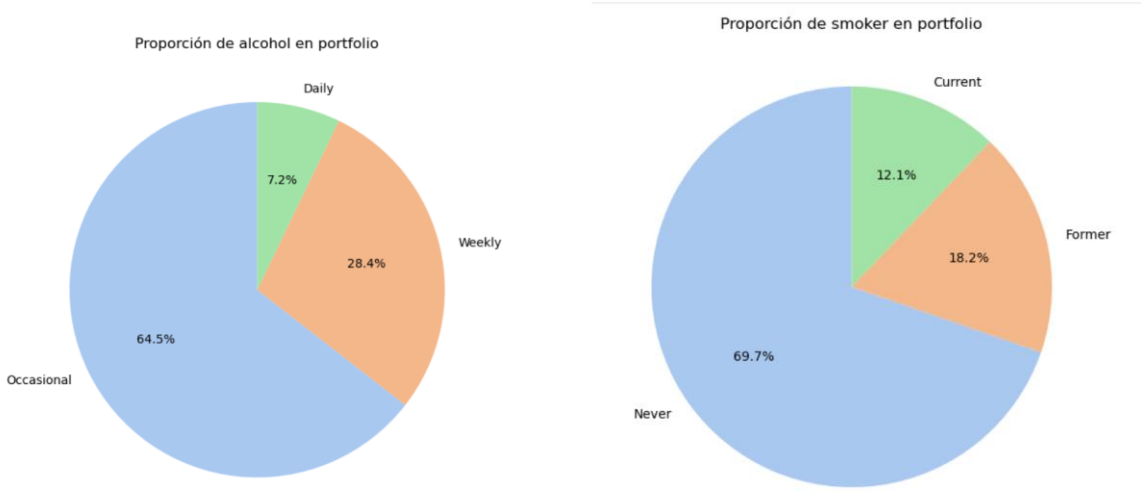


Anteriormente hemos visto que el primer factor no relacionado con el precio directamente que afecta el que mas al coste total es el riesgo; por lo que vamos a ver un poco como evoluciona el riesgo según diferentes factores. Aquí se puede ver como claramente los puntos verdes (Current o former tienen mucho mas riesgo) con la misma edad; que aquellos que no fuman (morado oscuro; por lo que e sabemos que ser fumador (actual afecta mucho al riesgo); luego ser “former” algo, y que si no eres fumador en general tienes mucho menos riesgo

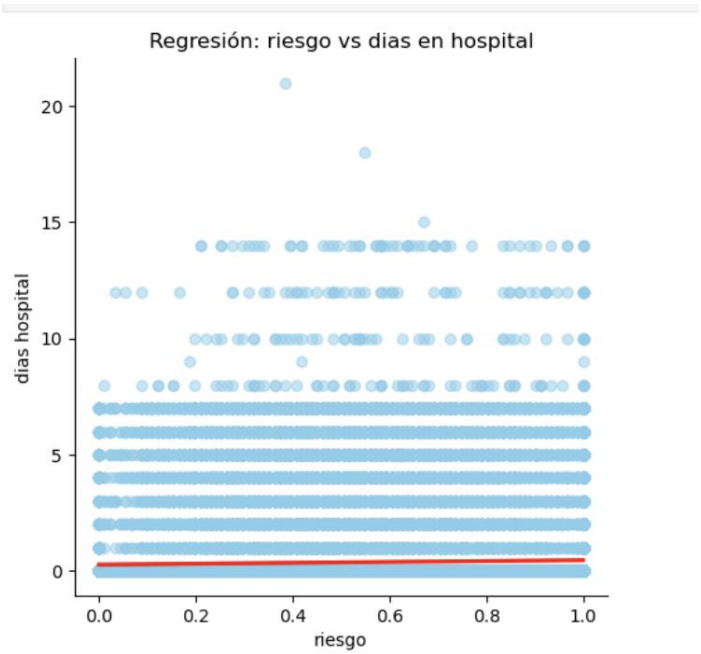


también vemos como la región no afecta al riesgo medio, será una variables que no tendremos que tener tanto en cuenta

Finalmente suponemos que beber alcohol y fumar afectara a nuestros costes; pero para poder visualizar esto primero necesitaremos saber el porcentaje que hay de cada uno en total:



En ambos casos vemos como no está estabilizado, por lo que esto lo hemos tenido en cuenta a la hora de hacer dashboard (los usaremos con filtro ya que sino, al tener mucho mas peso en los que no fuman, o que casi no beben esto puede afectar mucho.



Finalmente vemos como dos variables muy importantes (riesgo y dias en el hospital, no tienen una relación lineal (por lo que las incluimos las dos en el modelo posteriormente sin gran riesgo a colinealidad

3. Modelo predictivo explicado y con tablas

Escribe aquí sobre tu modelo, qué predice, qué variables ha utilizado, métricas o tablas de evaluación y gráfico de coeficientes para explicar el modelo.
He usado un modelo de regresión lineal para predecir el coste total. Priemro usando todas las variables numéricas (quitando aquellas con correlación > 0,8 visto enlas correlaciones

anteriores) y luego quedándonos solo con 10 variables para intentar quitar “ruido”. En los dos casos nos sale un R2 muy bajo indicando que predice bastante mas el coste total.

Esto no se puede indicar que a pesar de haber muchas variables que afectan, se debería analizar mas profundamente para poder predecir mejor. Es posible que como mezclamos muchas categorías y muchas enfermedades, condiciones (fumar, beber) que no este pudiendo hacer esta buena predicción:

Modelo 1: todas las var:

	Métrica	Train	Test
0	R2	6.433654e-01	5.967600e-01
1	MSE	3.482899e+06	3.967941e+06
2	RMSE	1.866253e+03	1.991969e+03
3	MAE	1.132398e+03	1.161999e+03

```
=== COEFICIENTES DE LA REGRESIÓN LINEAL ===
                     Variable  Coeficiente
19                  risk_score  2175.616776
8   days_hospitalized_last_3yrs   227.159313
23                chronic_count   127.182092
7    hospitalizations_last_3yrs    86.136548
40          had_major_procedure    55.766191
29               cancer_history    53.113780
25                     diabetes    37.550634
24                 hypertension    28.697847
17    policy_changes_last_2yrs    28.625073
28      cardiovascular_disease    23.620475
31                liver_disease    21.787513
32                    arthritis    11.616167
37           proc_consult_count    11.508140
4                    dependents     4.431850
5                           bmi     3.859246
18            provider_quality     3.563552
34           proc_imaging_count     1.282442
22            total_claims_paid     1.053953
16            policy_term_years     0.740905
12                          ldl     0.236210
21             avg_claim_amount     0.166431
14                   deductible     0.003005
0                     person_id    -0.000053
2                        income    -0.000240
38              proc_lab_count    -0.214616
15                       copay    -0.226785
11                 diastolic_bp    -0.379495
10                  systolic_bp    -1.333762
36           proc_physio_count    -1.799761
26                      asthma    -4.090550
9              medication_count    -8.188738
3                household_size    -9.031947
30               kidney_disease    -9.048578
1                           age   -10.866643
13                        hba1c   -13.872422
27                         copd   -17.025380
33                mental_health   -19.039817
6              visits_last_year   -36.794511
35           proc_surgery_count   -44.864413
39                 is_high_risk   -74.531289
20                 claims_count  -420.933994
```

Modelo 2: Solo algunas (peor modelo)

:

	Métrica	Train	Test
0	R2	1.576806e-01	1.661885e-01
1	MSE	3.482899e+06	3.967941e+06
2	RMSE	1.866253e+03	1.991969e+03
3	MAE	1.132398e+03	1.161999e+03

```
=== COEFICIENTES DE LA REGRESIÓN LINEAL ===
                      Variable  Coeficiente
0                   risk_score  4001.247464
2   days_hospitalized_last_3yrs   483.014150
1                chronic_count   330.190004
6                 hypertension    73.888740
9                    arthritis    45.893488
11                    diabetes     8.810319
5                          bmi     6.333990
7        cardiovascular_disease     3.021596
12                      income    -0.000432
8                  systolic_bp    -1.731274
10            provider_quality    -8.626273
3             medication_count   -11.840297
4                          age   -21.257314
```