Relatório EDA

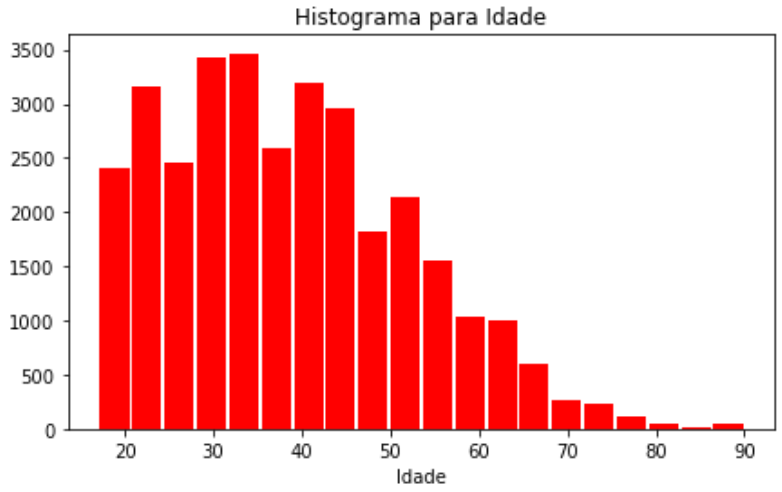
Por Bernardo Cesar C. P. Toazza

Proposto o problema de prever se os indivíduos de um dataframe receberiam salários inferiores ou superiores a 50k anualmente, precisamos dividir o problema em partes para facilitar sua solução. Como já temos uma base de dados pronta, sem problemas de tratamento, faremos a análise exploratória desses dados fornecidos com o objetivo de compreendê-los.

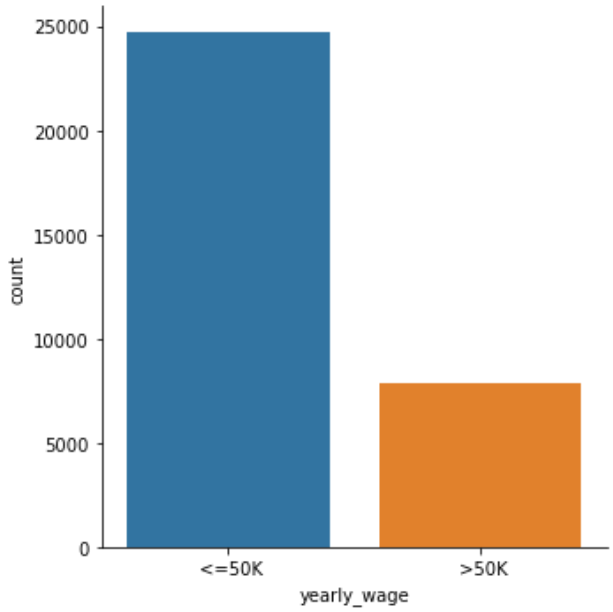
Estatística descritiva básica:

	age	fnlwgt	education_num	capital_gain	capital_loss	hours_per_week
count	32560.000000	3.256000e+04	32560.000000	32560.000000	32560.000000	32560.000000
mean	38.581634	1.897818e+05	10.080590	1077.615172	87.306511	40.437469
std	13.640642	1.055498e+05	2.572709	7385.402999	402.966116	12.347618
min	17.000000	1.228500e+04	1.000000	0.000000	0.000000	1.000000
25%	28.000000	1.178315e+05	9.000000	0.000000	0.000000	40.000000
50%	37.000000	1.783630e+05	10.000000	0.000000	0.000000	40.000000
75%	48.000000	2.370545e+05	12.000000	0.000000	0.000000	45.000000
max	90.000000	1.484705e+06	16.000000	99999.000000	4356.000000	99.000000

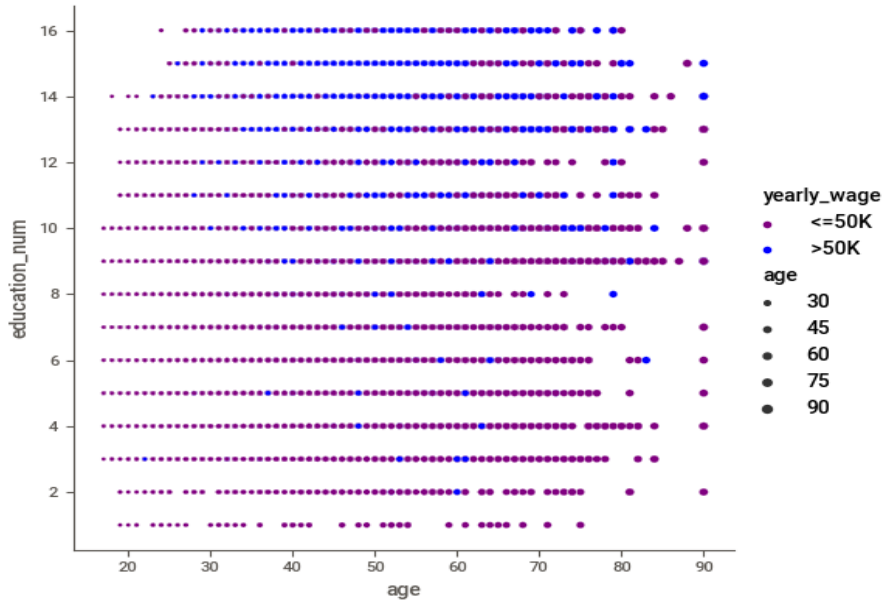
Qual a quantidade de pessoas por idade? Como se distribui? Com o histograma abaixo podemos perceber que há uma concentração de pessoas mais novas (até 45 anos) e menos nas demais idades (média de 38 anos).

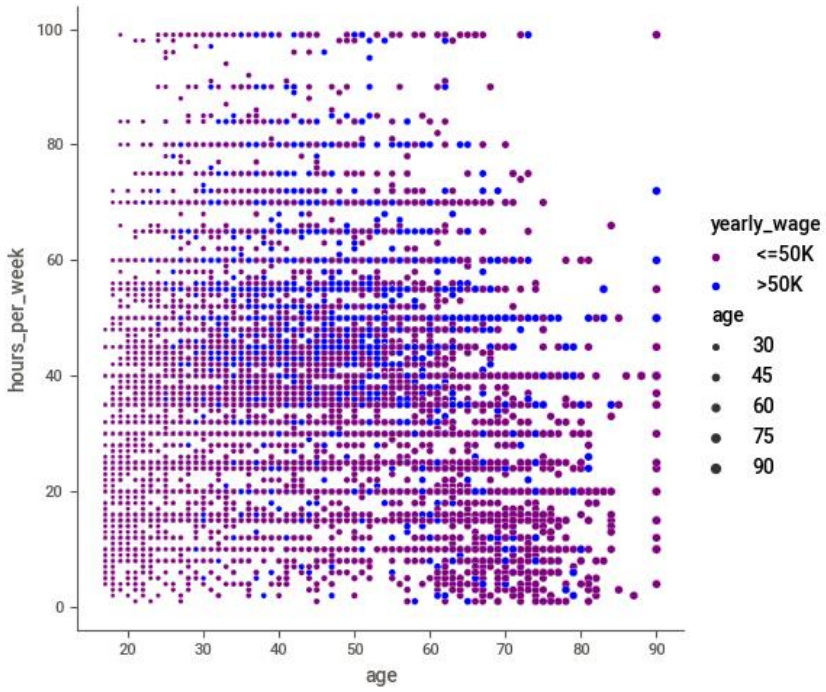


Há menos ou mais pessoas ganhando acima de 50K/ano? Ao fazer esse gráfico de contagem, fica evidente que há muito mais pessoas ganhando abaixo desse valor. Mas por quê?



Ao fazer um gráfico associando a idade das pessoas e tempo de educação, percebemos algo curioso: quanto mais velha e quanto mais "estudada", maior a tendência desta ganhar acima de 50K/ano. Isso nos ajuda a entender porque há poucas pessoas ganhando quantias maiores: há menos pessoas mais velhas (acima mostrado) e com mais tempo de estudo.

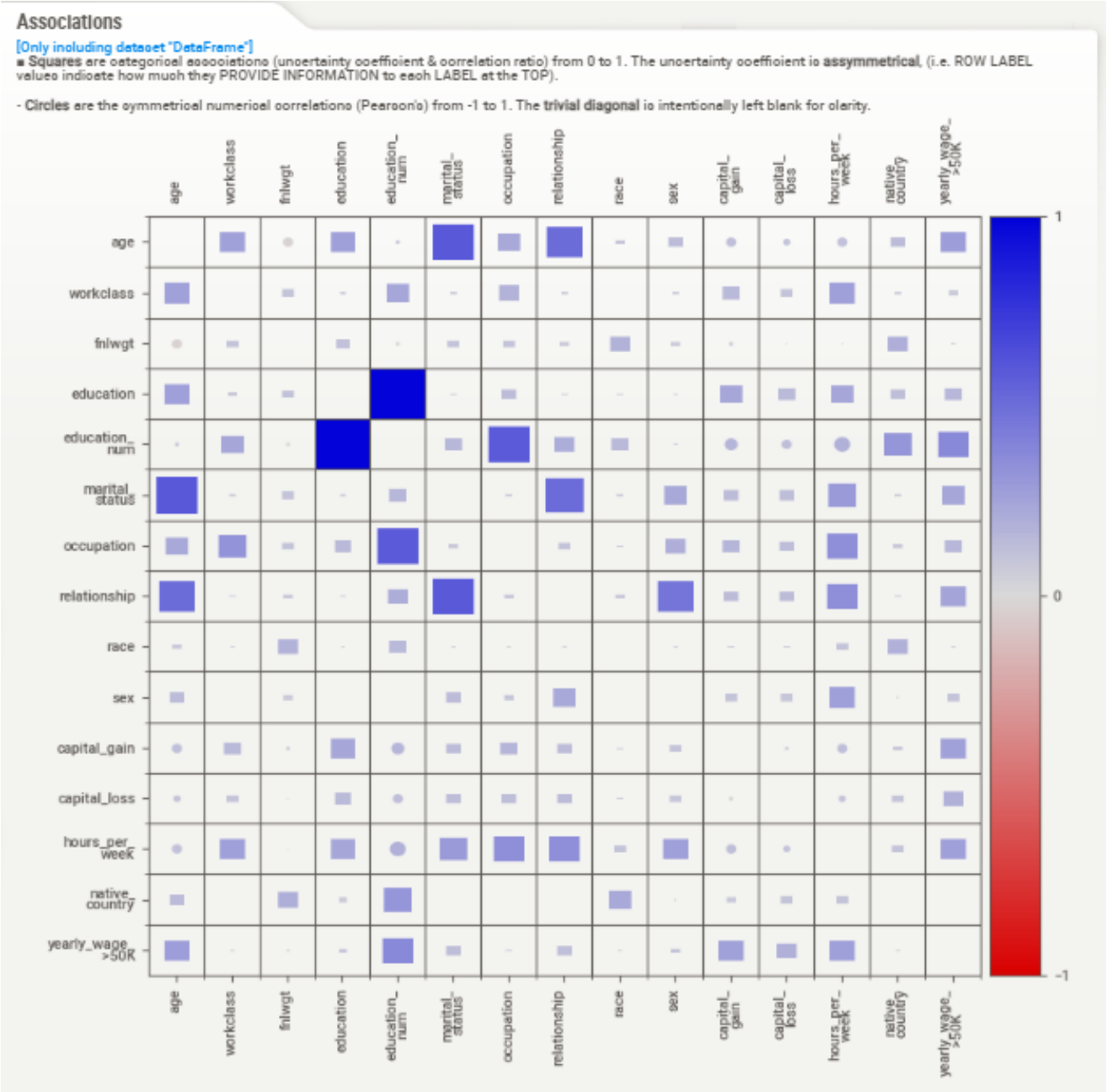




Com o gráfico acima percebemos que quanto mais idade e horasa trabalhadas semanalmente, maior os salários.

Associações

O pacote SweetViz nos fornece uma matriz de associação entre as variáveis muito interessante, sendo aqui filtrado as features que colocarei no modelo preditivo.



O modelo

A partir da análise exploratória dos dados, passemos ao modelo: do que se trata: uma classificação ou regressão. Por "yearly_wage" ser uma variável binária, ou seja, podendo ser maior do que 50K/ano ou inferior ou igual a isso, se trata de uma classificação, sendo o segundo indicado para prever variáveis contínuas (o que não é o caso). O lado bom de escolhe-lo é que é preciso e possui um bom desempenho em muitos problemas, inclusive não lineares. A

desvantagem é a falta de interpretabilidade e pode ocorrer overfitting com mais facilidade. A medida de performance foi a acurácia: acertos ao todo, o qual tem fica em 0.84, um número bom; e, f1 Score, que mede o nível de falsos positivos e negativos, que ficou em 0.65. Usei ela pois nos dá uma visão geral dos resultados, o quanto o modelo classificou corretamente.