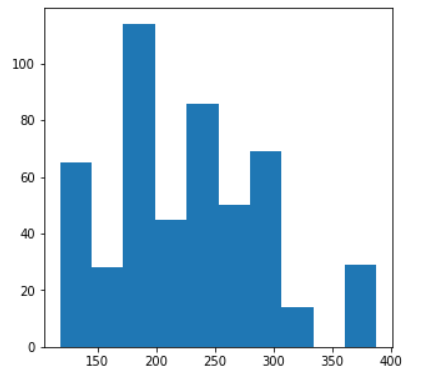
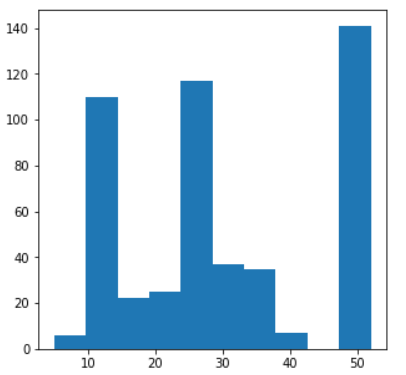
Estudo de Outliers

# Transportation Expenses



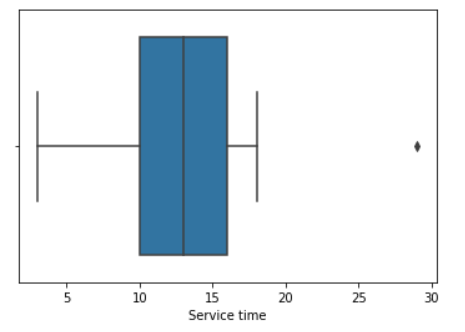
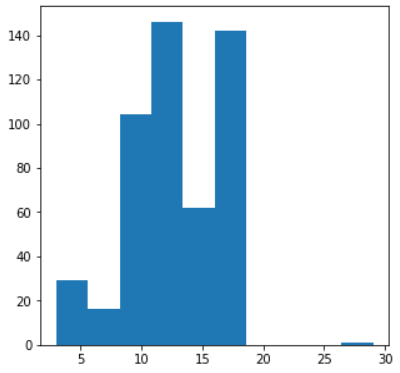
Não tem outliers.

# Distance from Residence to Work

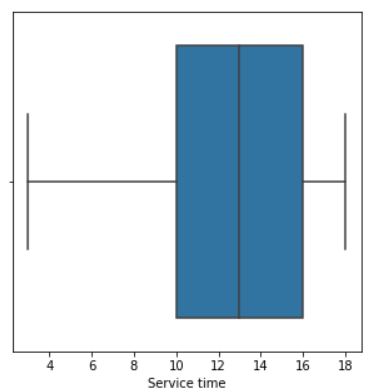
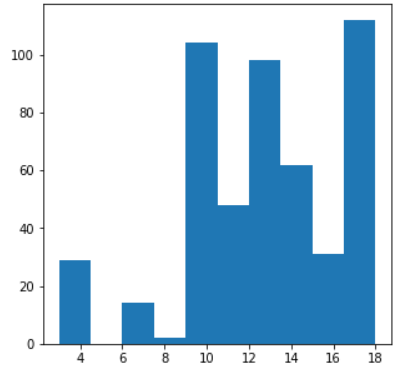


Não tem outliers.

# Service Time

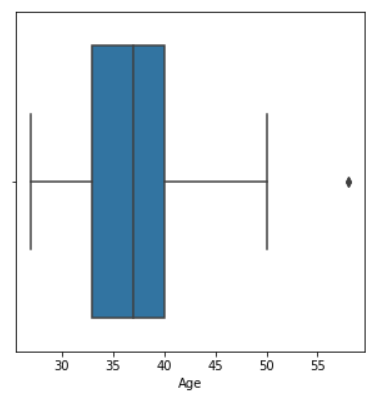
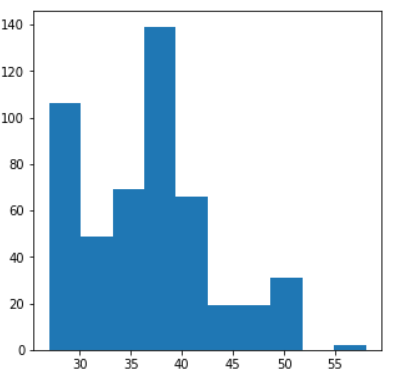


Tem um outlier óbvio. Vai ser passado ao máximo de service time.



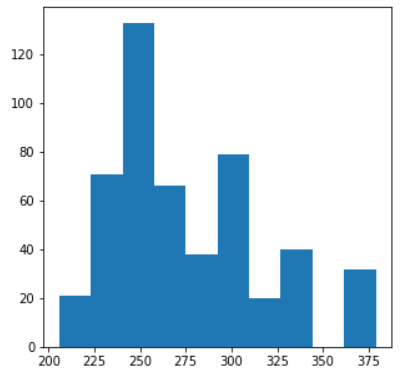
Distribuição após resolução.

# Age



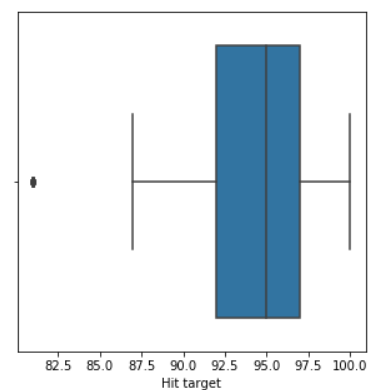
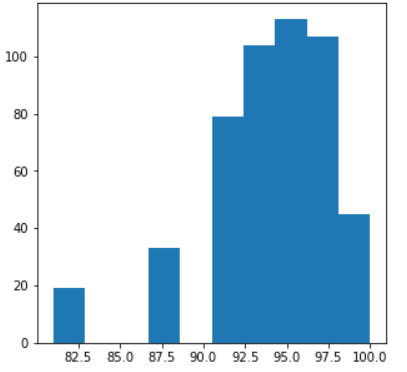
Não pode ser considerado outlier. De qualquer modo vai ser absorvido quando se fizer binning.

Work load Average/day

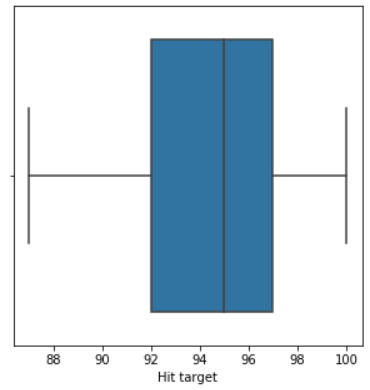
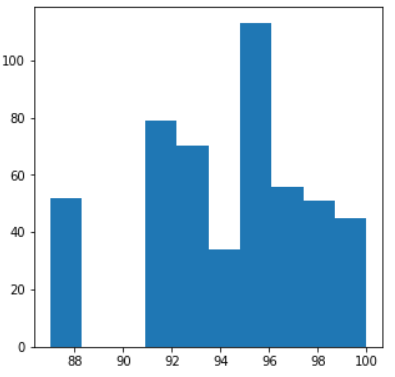


Não tem outliers.

# Hit target

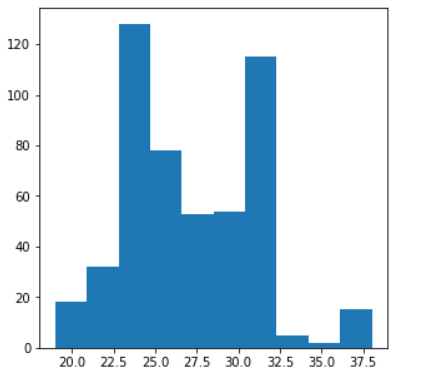


É outlier. Aqueles pontos vão passar a ter o valor do mínimo encontrado no percentil(0.05).



Outlier resolvido. O valor mínimo daqueles pontos passou a ser o mínimo do percentil 0.04 (correspondente ao mínimo da última barra).

Body mass índex



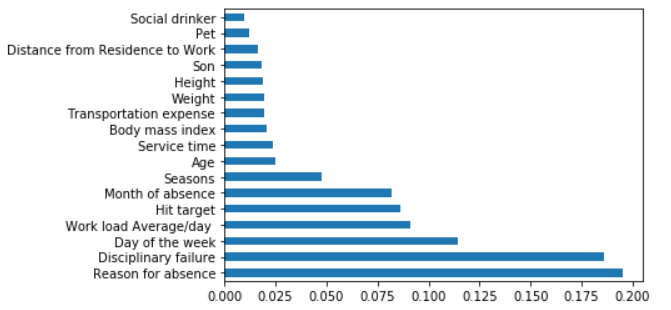
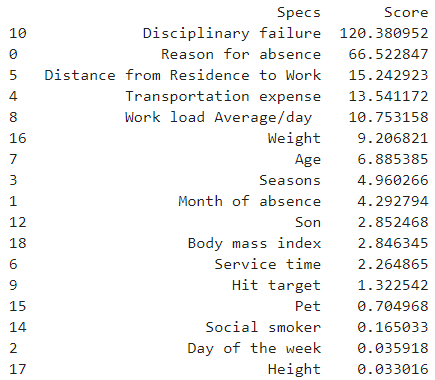
Não tem outliers.

Binning

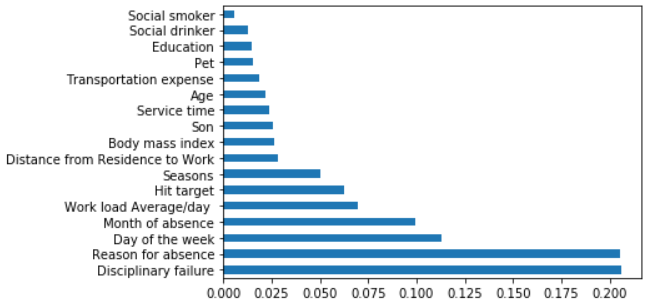
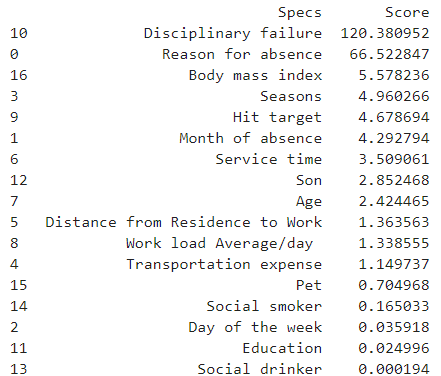
Feature Selection

# SelectKBest

Antes de remoção de outliers e binning.



Após a remoção de outliers e binning.



# One Hot Encoding

Havia mais valores diferentes para YES do que para NO. É possível verificar, desde logo, a importância desta variável.

O One Hot Encoding teve que ser feito com cuidado porque, devido ao pequeno tamanho do *dataset*, nem todos os valores possíveis para esta label apareciam no train e no test. Por este motivo, foi feita uma operação de limpeza e criação de colunas em ambos os subsets, por forma a uniformiza-los e ser possível passar ao modelo.

