Auxílio a Detecção de Fraude Imobiliária com Machine Learning

Utilizando inteligência artificial para prevenir fraudes no setor imobiliário

Emerson Silva, Cientista de Dados

O Problema

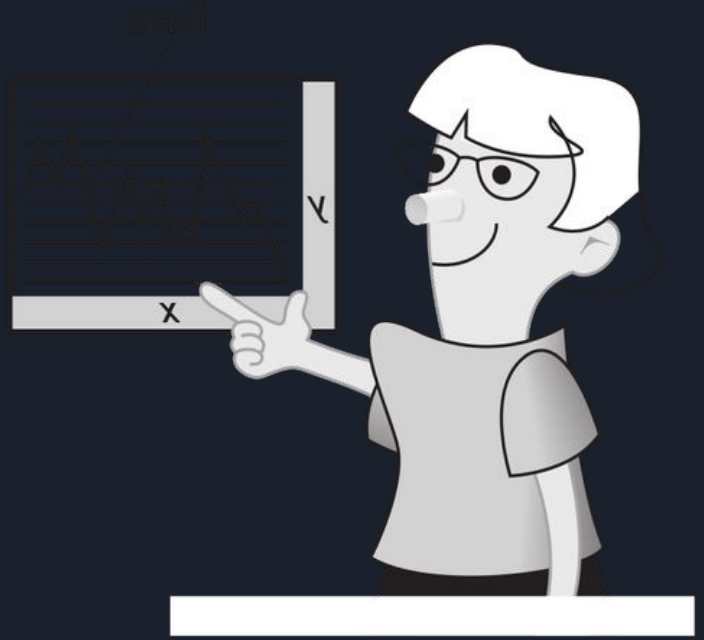
- Milhões de transações imobiliárias ocorrem anualmente
- Transações das mais variadas
- Algumas dessas transações podem ser fraudulentas
- Grandes perdas financeiras
- Impactos no mercado e questões legais
- Como identificar padrões suspeitos e prevenir fraudes?

Nossa Solução

- Explorar a base de dados buscando corrigir possíveis falhas para o trabalho
- Utilizar Machine Learning para análise e detecção de padrões suspeitos
- Criar um sistema inteligente para ranquear transações com maior risco de fraude
- Apresentar os dados organizados a partir das transações com maior risco de fraude para menor risco de fraude.

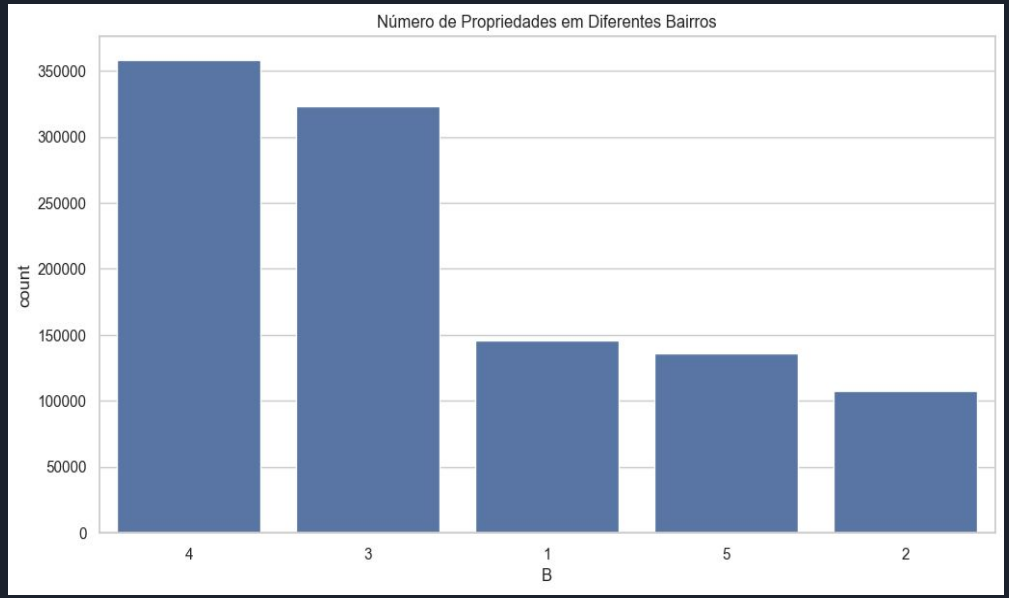
Entendendo os Dados

- Exploração do dataset
- Análise da Qualidade dos Dados (DQR)
- Verificação de valores ausentes
- Verificação outliers
- Inconsistências nos dados



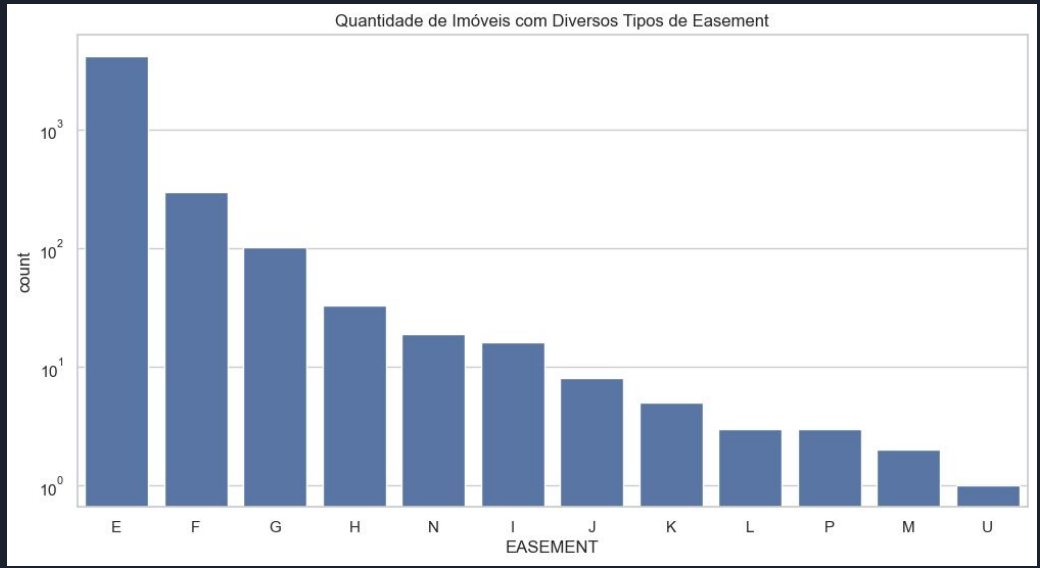
Entendendo os Dados

- Análise da Qualidade dos Dados (DQR)
  - Número de Propriedades em Diferentes Bairros



Entendendo os Dados

- Análise da Qualidade dos Dados (DQR)
  - Quantidade de Imóveis com Diversos Tipos de Easement



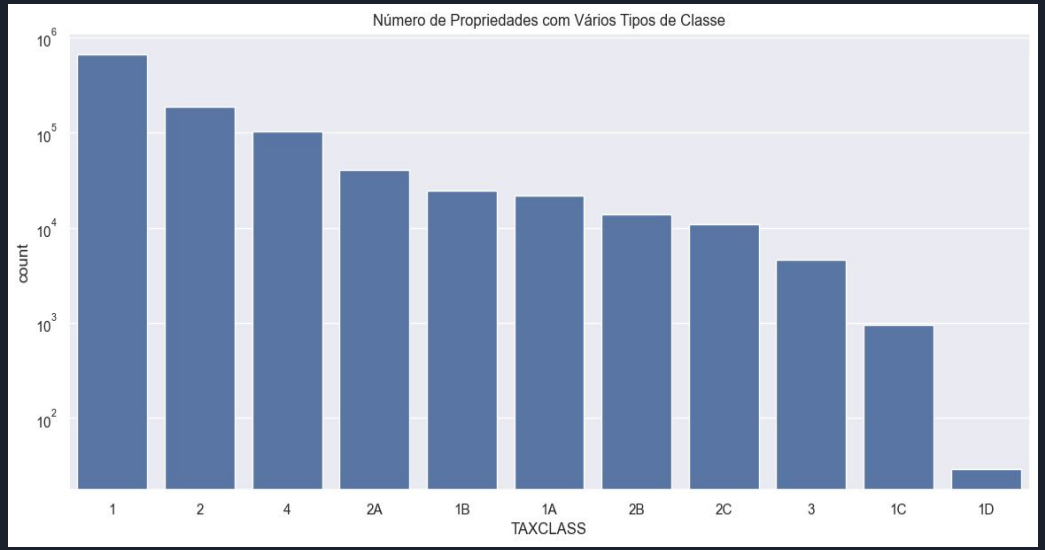
Entendendo os Dados

- Análise da Qualidade dos Dados (DQR)
  - Quantidade de Imóveis por Proprietário

	Unique_values_OWNER	Counts
0	PARKCHESTER PRESERVAT	6020
1	PARKS AND RECREATION	4255
2	DCAS	2169
3	HOUSING PRESERVATION	1904
4	CITY OF NEW YORK	1450
5	DEPT OF ENVIRONMENTAL	1166
6	BOARD OF EDUCATION	1015
7	NEW YORK CITY HOUSING	1014
8	CNY/NYCTA	975
9	NYC HOUSING PARTNERSH	747
10	YORKVILLE TOWERS ASSO	558
11	DEPARTMENT OF BUSINES	527
12	DEPT OF TRANSPORTATIO	503
13	MTA/LIRR	467
14	PARCKHESTER PRESERVAT	439
15	MH RESIDENTIAL 1, LLC	411
16	434 M LLC	393
17	LINCOLN PLAZA ASSOCIA	366
18	DEUTSCHE BANK NATIONA	336
19	561 11TH AVENUE TMG L	324

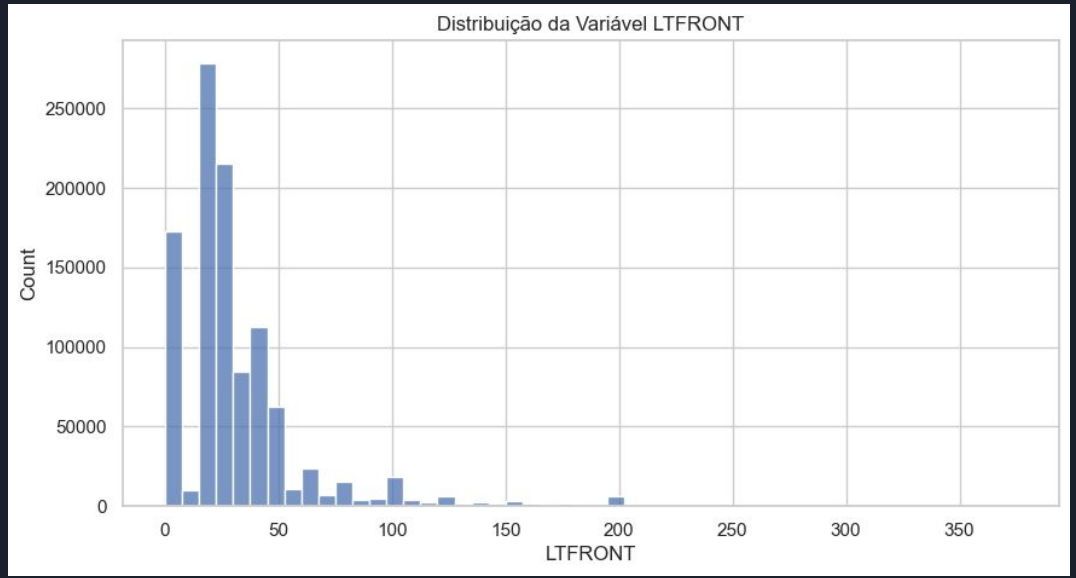
Entendendo os Dados

- Análise da Qualidade dos Dados (DQR)
  - Número de Propriedades por Classe de Imposto



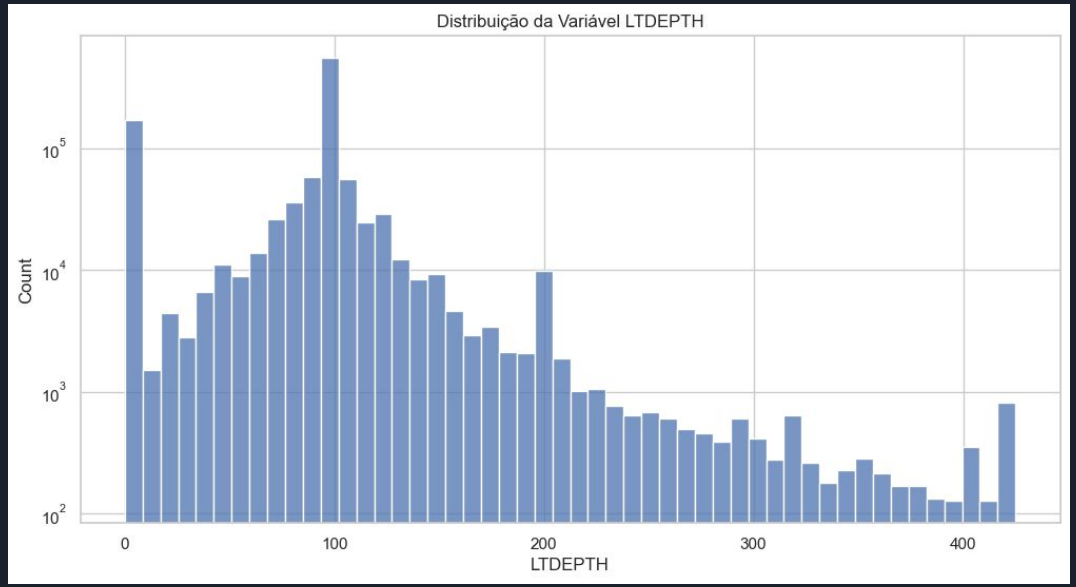
Entendendo os Dados

- Análise da Qualidade dos Dados (DQR)
  - Número de Propriedades por Tamanho de Frente do Lote em pés, filtrados para <= 375



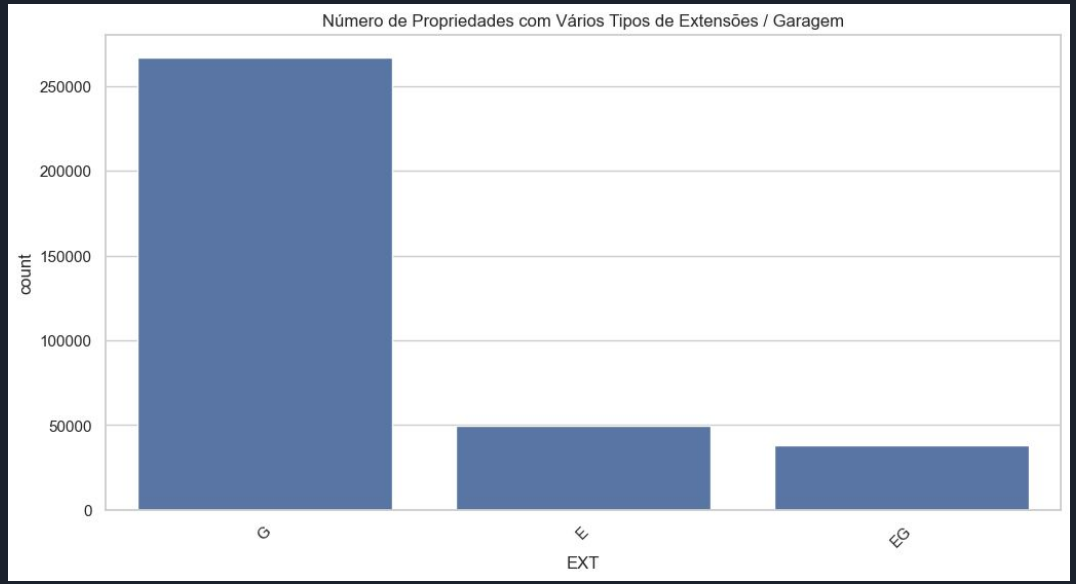
Entendendo os Dados

- Análise da Qualidade dos Dados (DQR)
  - Número de Propriedades por Profundidade do Lote em pés, filtrados para <= 425



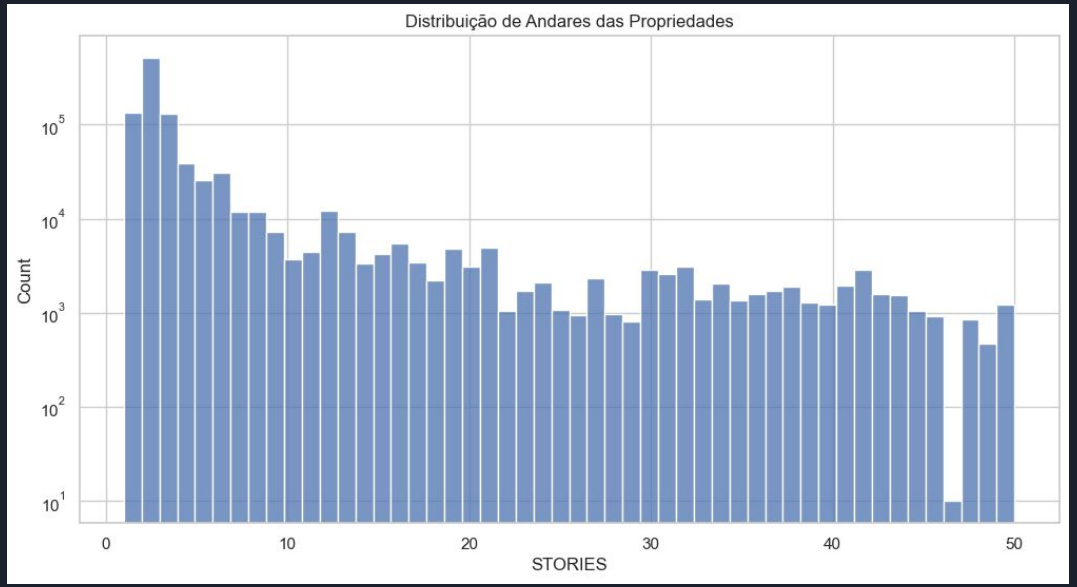
Entendendo os Dados

- Análise da Qualidade dos Dados (DQR)
  - Número de Propriedades por E- Extension, G- Garage, EG- Extension e Garage



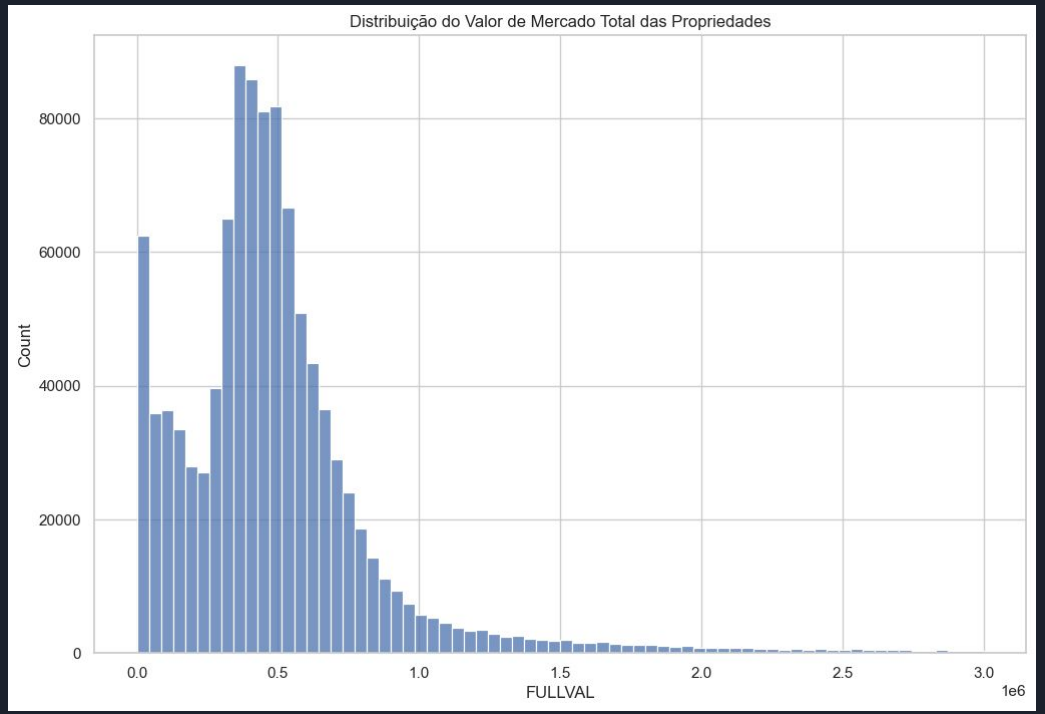
Entendendo os Dados

- Análise da Qualidade dos Dados (DQR)
  - Número de Propriedades por Número de Andares do Edifício, filtrados para <= 50



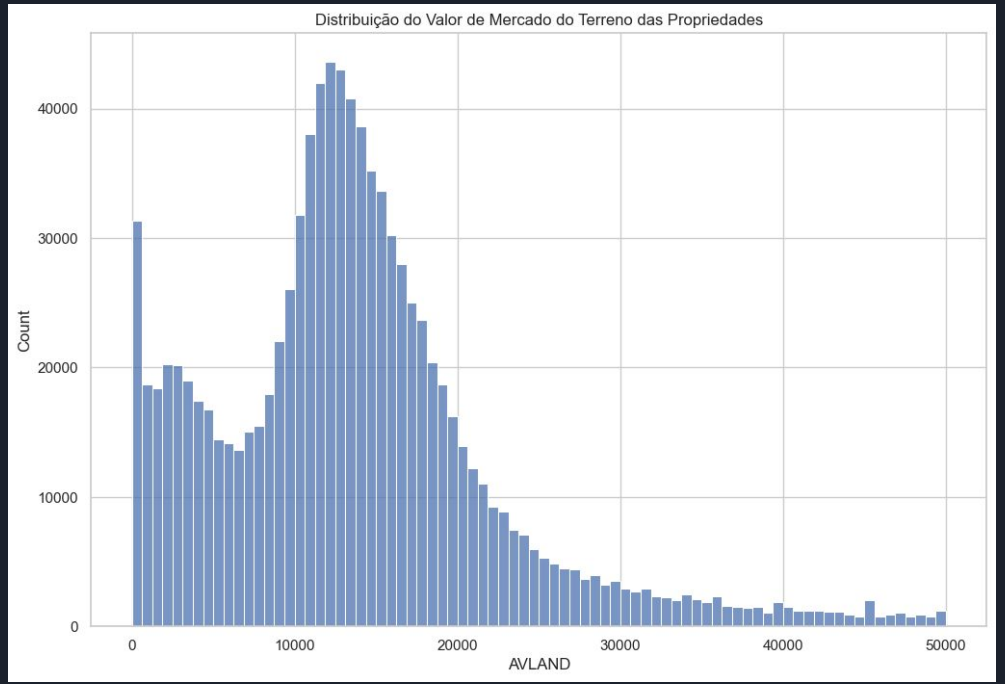
Entendendo os Dados

- Análise da Qualidade dos Dados (DQR)
  - Número de Propriedades por Valor de Mercado, filtrados para <= ($) 3M



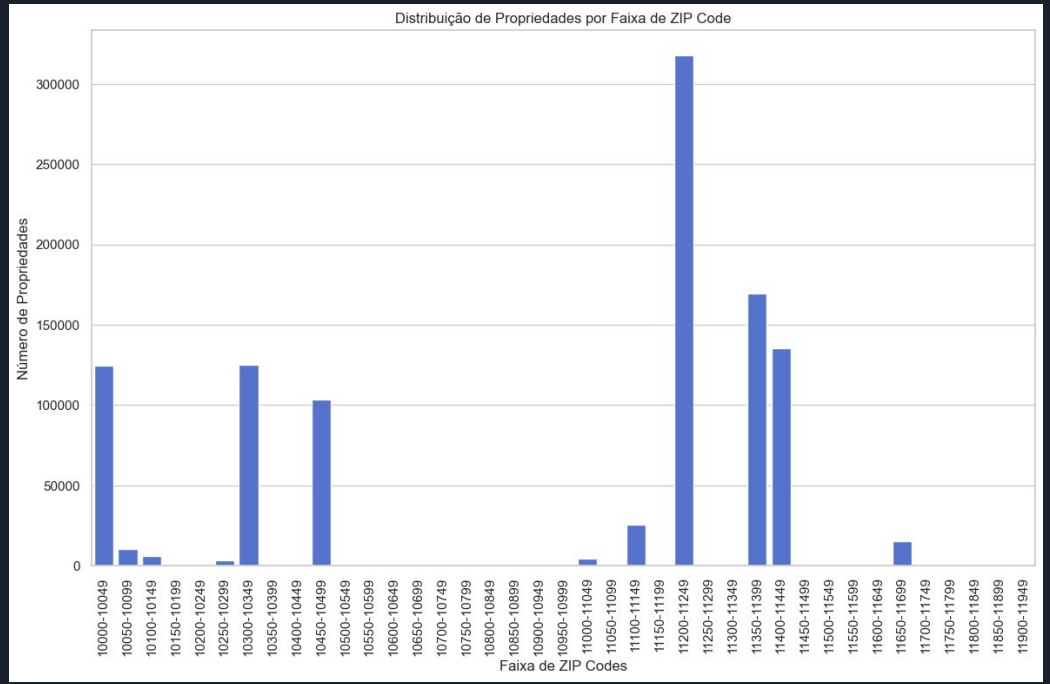
Entendendo os Dados

- Análise da Qualidade dos Dados (DQR)
  - Número de Propriedades por Valor de Mercado do Terreno, filtrados para <= ($) 50K



Entendendo os Dados

- Análise da Qualidade dos Dados (DQR)
  - Número de Propriedades por Faixa de ZIP Code



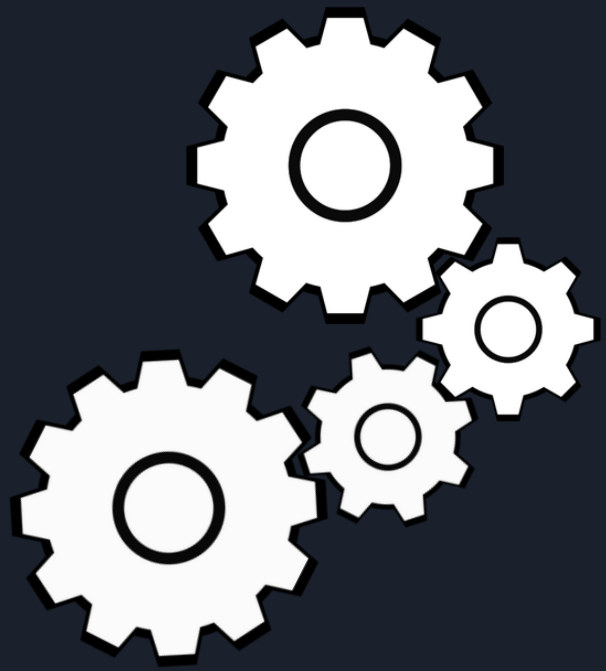
Tratando Valores Ausentes

- Registros ZIP ausentes através da primeira moda de cada grupo
- Registros vazios com a mediana dos grupos:
  - ZIP e BLDGCL
  - ZIP e TAXCLASS
  - B e TAXCLASS
  - B
  - BLDGCL
  - TAXCLASS

```
Valores Ausentes:
 RECORD             0
BBLE                0
B                   0
BLOCK               0
LOT                 0
EASEMENT      1066358
OWNER           31745
BLDGCL              0
TAXCLASS            0
LTFRONT             0
LTDEPTH             0
EXT            716689
STORIES         56264
FULLVAL             0
AVLAND              0
AVTOT               0
EXLAND              0
EXTOT               0
EXCD1          432506
STADDR            676
ZIP             29890
EXMPTCL       1055415
BLDFRONT            0
BLDDEPTH            0
AVLAND2        788268
AVTOT2         788262
EXLAND2        983545
EXTOT2         940166
EXCD2          978046
PERIOD              0
YEAR                0
VALTYPE             0
```

Engenharia de Atributos - Extraindo Informações dos Dados

- Área 1 = LTFRONT * LTDEPTH
- Área 2 = BLDFRONT * BLDDEPTH
- Área 3 = AREA2 * STORIES
- Gerando um Índice através da divisão de FULLVAL, AVLAND e AVTOT pelas variáveis recém criadas AREA1, AREA2 e AREA3
  - ind1 é a combinação feita entre: FULLVAL AREA1
  - ind2 é a combinação feita entre: FULLVAL AREA2
  - ind3 é a combinação feita entre: FULLVAL AREA3
  - ind4 é a combinação feita entre: AVLAND AREA1
  - ind5 é a combinação feita entre: AVLAND AREA2
  - ind6 é a combinação feita entre: AVLAND AREA3
  - ind7 é a combinação feita entre: AVTOT AREA1
  - ind8 é a combinação feita entre: AVTOT AREA2
  - ind9 é a combinação feita entre: AVTOT AREA3



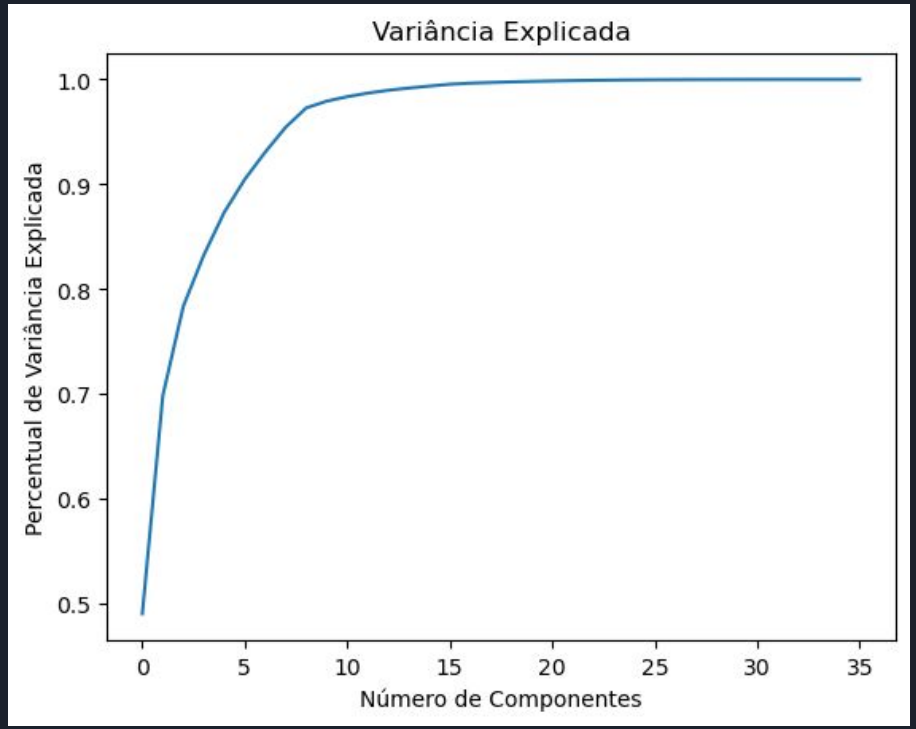
Engenharia de Atributos - Extraindo Informações dos Dados

- Variáveis originais e pós engenharia de atributos

```
['RECORD', 'BBLE', 'B', 'BLOCK', 'LOT', 'EASEMENT', 'OWNER', 'BLDGCL',
 'TAXCLASS', 'LTFRONT', 'LTDEPTH', 'EXT', 'STORIES', 'FULLVAL', 'AVLAND',
 'AVTOT', 'EXLAND', 'EXTOT', 'EXCD1', 'STADDR', 'ZIP', 'EXMPTCL',
 'BLDFRONT', 'BLDDEPTH', 'AVLAND2', 'AVTOT2', 'EXLAND2', 'EXTOT2',
 'EXCD2', 'PERIOD', 'YEAR', 'VALTYPE', 'ind1_media_ind1_grupo_ZIP',
 'ind2_media_ind2_grupo_ZIP', 'ind3_media_ind3_grupo_ZIP',
 'ind4_media_ind4_grupo_ZIP', 'ind5_media_ind5_grupo_ZIP',
 'ind6_media_ind6_grupo_ZIP', 'ind7_media_ind7_grupo_ZIP',
 'ind8_media_ind8_grupo_ZIP', 'ind9_media_ind9_grupo_ZIP',
 'ind1_media_ind1_grupo_TAXCLASS', 'ind2_media_ind2_grupo_TAXCLASS',
 'ind3_media_ind3_grupo_TAXCLASS', 'ind4_media_ind4_grupo_TAXCLASS',
 'ind5_media_ind5_grupo_TAXCLASS', 'ind6_media_ind6_grupo_TAXCLASS',
 'ind7_media_ind7_grupo_TAXCLASS', 'ind8_media_ind8_grupo_TAXCLASS',
 'ind9_media_ind9_grupo_TAXCLASS', 'ind1_media_ind1_grupo_B',
 'ind2_media_ind2_grupo_B', 'ind3_media_ind3_grupo_B',
 'ind4_media_ind4_grupo_B', 'ind5_media_ind5_grupo_B',
 'ind6_media_ind6_grupo_B', 'ind7_media_ind7_grupo_B',
 'ind8_media_ind8_grupo_B', 'ind9_media_ind9_grupo_B',
 'ind1_media_ind1_grupo_All', 'ind2_media_ind2_grupo_All',
 'ind3_media_ind3_grupo_All', 'ind4_media_ind4_grupo_All',
 'ind5_media_ind5_grupo_All', 'ind6_media_ind6_grupo_All',
 'ind7_media_ind7_grupo_All', 'ind8_media_ind8_grupo_All',
 'ind9_media_ind9_grupo_All'],
```

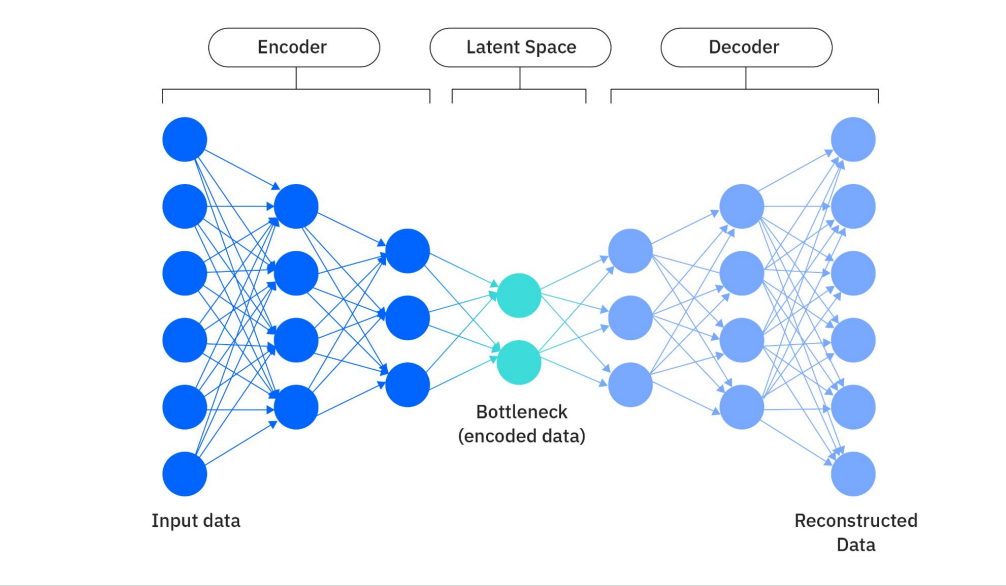
Como Identificamos Fraudes?

- PCA com 15 Componentes
  - São criadas 15 scores, onde cada score é o quadrado do valor da componente principal correspondente
  - Soma os 15 scores para cada linha e tira a raiz quadrada
  - Quanto maior o valor, mais anômalo (chance de fraude) é o registro



Machine Learning

- Autoencoder (Deep Learning)
  - Rede neural que aprende uma representação comprimida dos dados (encoding) e tenta reconstruí-los (decoding)
  - Se houver um grande erro de reconstrução, significa que a entrada pode ser uma anomalia
  - O score é a raiz quadrada da soma dos erros, ou seja, quanto maior o score, mais anômala a amostra (chance de fraude)



O Ranking de Fraude

- Calculamos um score final ponderado, combinando os dois scores
  - Ordenando os valores do maior para o menor
  - Quanto maior o score maior chance de fraude

_All	ind7_media_ind7_grupo_All	ind8_media_ind8_grupo_All	ind9_media_ind9_grupo_All	Fraud Score 1	Fraud Score 2	Rank_Fraud Score 1	Rank_Fraud Score 2	Final Score	Final Rank
585	2.887832	17366.599665	38021.252085	1032.370884	1024.639358	1070994.0	1070994.0	2.203046e+09	1.0
540	11692.901316	21.429852	46.917061	1021.876097	1019.530897	1070993.0	1070993.0	2.186333e+09	2.0
905	5033.093327	24656.466391	26990.595231	925.736545	925.736545	1070992.0	1070992.0	1.982913e+09	3.0
437	1638.598722	0.407828	0.446436	916.675186	916.675186	1070991.0	1070991.0	1.963502e+09	4.0
542	310.594086	29280.261695	21368.071135	902.674789	902.674789	1070990.0	1070990.0	1.933511e+09	5.0
264	1.163778	2815.404383	1027.309143	811.364025	801.377377	1070989.0	1070989.0	1.727228e+09	6.0
527	0.422545	20409.607265	2234.170297	775.441577	775.441577	1070988.0	1070988.0	1.660977e+09	7.0
438	3.518226	22001.644354	4816.890367	740.377257	740.377257	1070987.0	1070987.0	1.585869e+09	8.0
779	0.851624	6164.253806	1686.949751	722.396934	713.516036	1070986.0	1070986.0	1.537843e+09	9.0
816	1.720284	6502.658004	14236.477143	638.775750	638.775750	1070985.0	1070985.0	1.368238e+09	10.0
907	72.248477	13170.189280	14416.958307	576.154995	571.717088	1070984.0	1070984.0	1.229353e+09	11.0
413	160.691131	0.077132	0.033773	493.699427	493.699427	1070983.0	1070983.0	1.057487e+09	12.0
290	0.665347	12752.892778	1396.015802	484.522409	484.522409	1070982.0	1070982.0	1.037830e+09	13.0
139	1142.426997	147.744635	323.462055	475.126561	475.126561	1070979.0	1070981.0	1.017702e+09	14.0

*** será disponibilizada tabela completa

O que ganhamos?

- Priorização nas investigações
- Ganho de tempo
- Economia de recursos
- Automação eficiente no setor imobiliário
- Maior segurança jurídica
- Maior segurança aos clientes