INSTITUTO FEDERAL
São Paulo
Câmpus Campinas

Projeto Interdisciplinar
Exploração e Aplicação de Técnicas de Ciência de Dados

Thales A. P. Pomari - CP 3013456

INSTITUTO FEDERAL
São Paulo
Câmpus Campinas

Roteiro

1. Base utilizada;

2. Pré-processamento aplicado;

3. Análise Exploratória;

4. Modelos de Regressão Logística;

5. Conclusões.

INSTITUTO FEDERAL
São Paulo
Câmpus Campinas

Base de Dados

- Brazil Highway Traffic Accidents;
- Ocorrências em rodovias federais;
- 2010 - 2015.

INSTITUTO FEDERAL
São Paulo
Câmpus Campinas

Base de Dados

- 26 colunas;
- Ausência de campos nulos inicialmente;
- Tipo das colunas incorretos.

Coluna ▲	Exemplo	Tipo
ano	2010	object
br	285	object
causa_acidente	Velocidade incompatível	object
classificacao_acidente	Sem Vítimas	object
condicao_meteorologica	Chuva	object
data_inversa	2010-10-29	object
dia_semana	Sexta	object
fase_dia	Pleno dia	object
feridos	0	int64
feridos_graves	0	int64
feridos_leves	0	int64
horario	14:20:00	object
id	1000329	int64
ignorados	0	int64

1 - 20 / 26 < >

INSTITUTO FEDERAL
São Paulo
Câmpus Campinas

Pré-Processamento e Limpeza

- Campo 'data_inversa';

- Padronização dos campos para o tipo date;

- Extração do mês.

ano	total_registros	yyyy-MM-dd	dd/MM/yyyy
2015	122161	122161	0
2014	169201	169201	0
2013	186748	186748	0
2012	184568	184568	0
2011	192326	0	192326
2010	183469	0	183469

INSTITUTO FEDERAL
São Paulo
Câmpus Campinas

Pré-Processamento e Limpeza

- Campo 'horario';
- Substituição pela hora;

horario	hora
22:00:00	22
16:00:00	16
14:20:00	14
12:30:00	12

INSTITUTO FEDERAL
São Paulo
Câmpus Campinas

Pré-Processamento e Limpeza

- Campo 'br';
- Tipos diferentes.

```
In [50]: df_ocorrencias_limpo.br.unique()

Out[50]: array([285, 116, 407, 101, 280, 381, 251, 369, 40, 232, 60, 316, 226, 174,
       277, 104, 262, 135, 70, 158, 267, 153, 463, 408, 10, 290, 282, 364,
       343, 110, 476, 317, 319, 259, 376, 324, 293, 163, 242, 392, 20,
       424, 230, 393, 480, 465, 365, 470, 222, 50, 386, 452, 354, 428,
       235, 287, 356, 471, 405, 367, 272, 373, 487, 459, 361, 402, 377,
       330, 423, 468, 427, 210, 304, 406, 493, 472, 450, 308, 146, 460,
       467, 414, 447, 410, 495, 156, 80, 469, 488, 418, 401, 474, 416,
       485, 432, 404, 425, 429, 419, 958, 30, 412, 490, 0, 409, 422, 208,
       400, 173, 359, 142, 1, 473, 420, 332, 870, 544, 298, 756, 707, 498,
       155, 421, '282', '280', '116', '364', '101', '319', '262', '40',
       '135', '459', '153', '381', '470', '50', '316', '110', '354',
       '393', '369', '324', '285', '70', '290', '277', '232', '158',
       '267', '242', '304', '392', '293', '386', '163', '104', '10',
       '235', '376', '230', '419', '365', '287', '174', '476', '343',
       '367', '60', '408', '427', '20', '414', '356', '469', '251', '493',
       '452', '424', '222', '146', '463', '406', '407', '428', '272',
       '480', '308', '226', '465', '450', '467', '373', '155', '317',
       '472', '377', '471', '423', '210', '405', '418', '259', '402',
       '80', '410', '412', '330', '361', '468', '634', '487', '488',
       '447', '401', '359', '156', '404', '495', '30', '474', '460',
       '415', '485', '429', '(null)', '349', '0', 4, 140, 28, 661, 462,
       349, 352, 617, 580, '432', '221', '560', '654', '241', '499',
       '420', '501', '489', '416', '473', '84', '687', '425', '178',
       '552', '453', '505', '183', 453, 265, 426, 270, 441, 152, 681, 154,
       767, 719, 323, 337, 851, 268, 2, 184, 415, 648, 380, 591, 448, 37,
       388, 884, 250, 931, 436, 433, 211, 417, 186], dtype=object)
```

INSTITUTO FEDERAL
São Paulo
Câmpus Campinas

Pré-Processamento e Limpeza

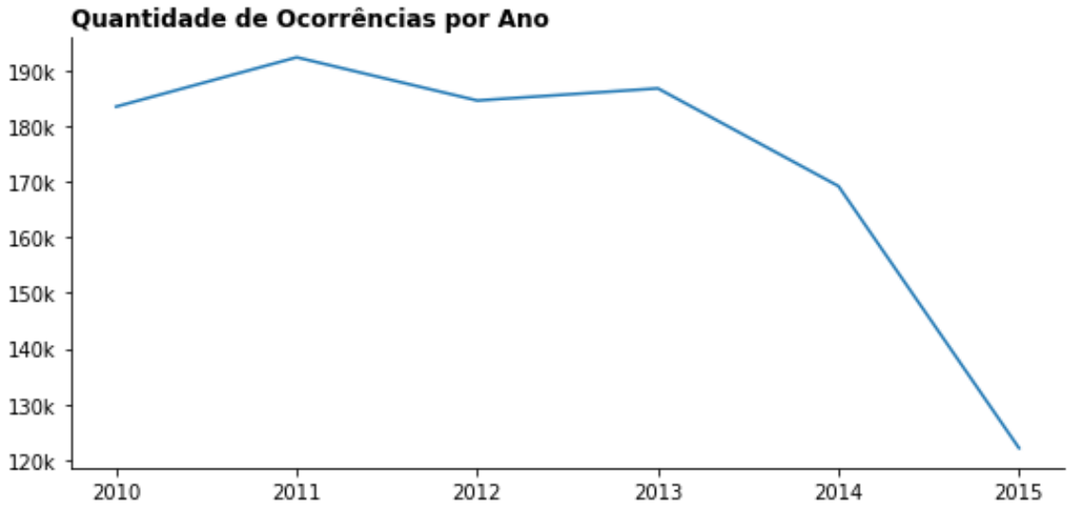
- Base limpa;
- '(null)' apareceu mais vezes.

coluna	tipo	exemplo
veiculos	int64	1
uso_solo	category	Rural
uf	category	RS
tracado_via	category	Curva
tipo_pista	category	Simples
tipo_acidente	category	Saída de Pista
sentido_via	category	Crescente
pessoas	int64	5
municipio	category	santa barbara d...
mortos	int64	0
mes	category	10
km	float64	397.3
ilesos	int64	5

1 - 25 / 27

INSTITUTO FEDERAL
São Paulo
Câmpus Campinas

Análise Exploratória

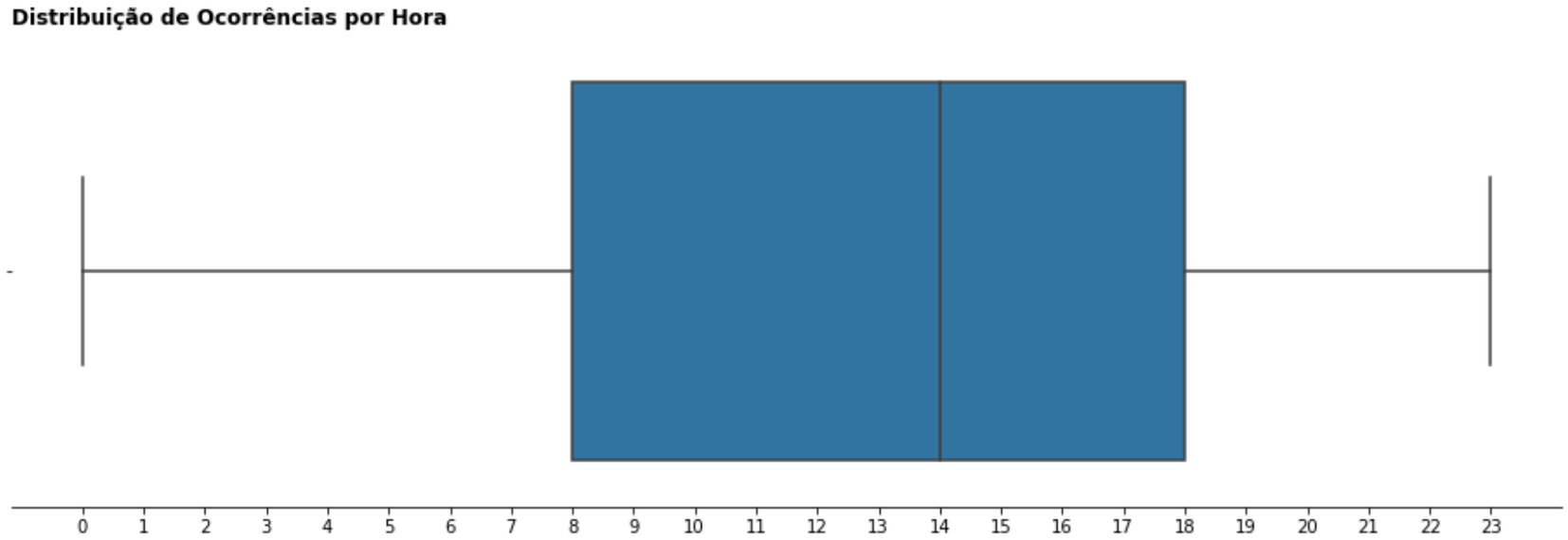


Como está a distribuição das ocorrências ao longo dos anos?

Qual é o mês com maior quantidade de acidentes acumulados?

INSTITUTO FEDERAL
São Paulo
Câmpus Campinas

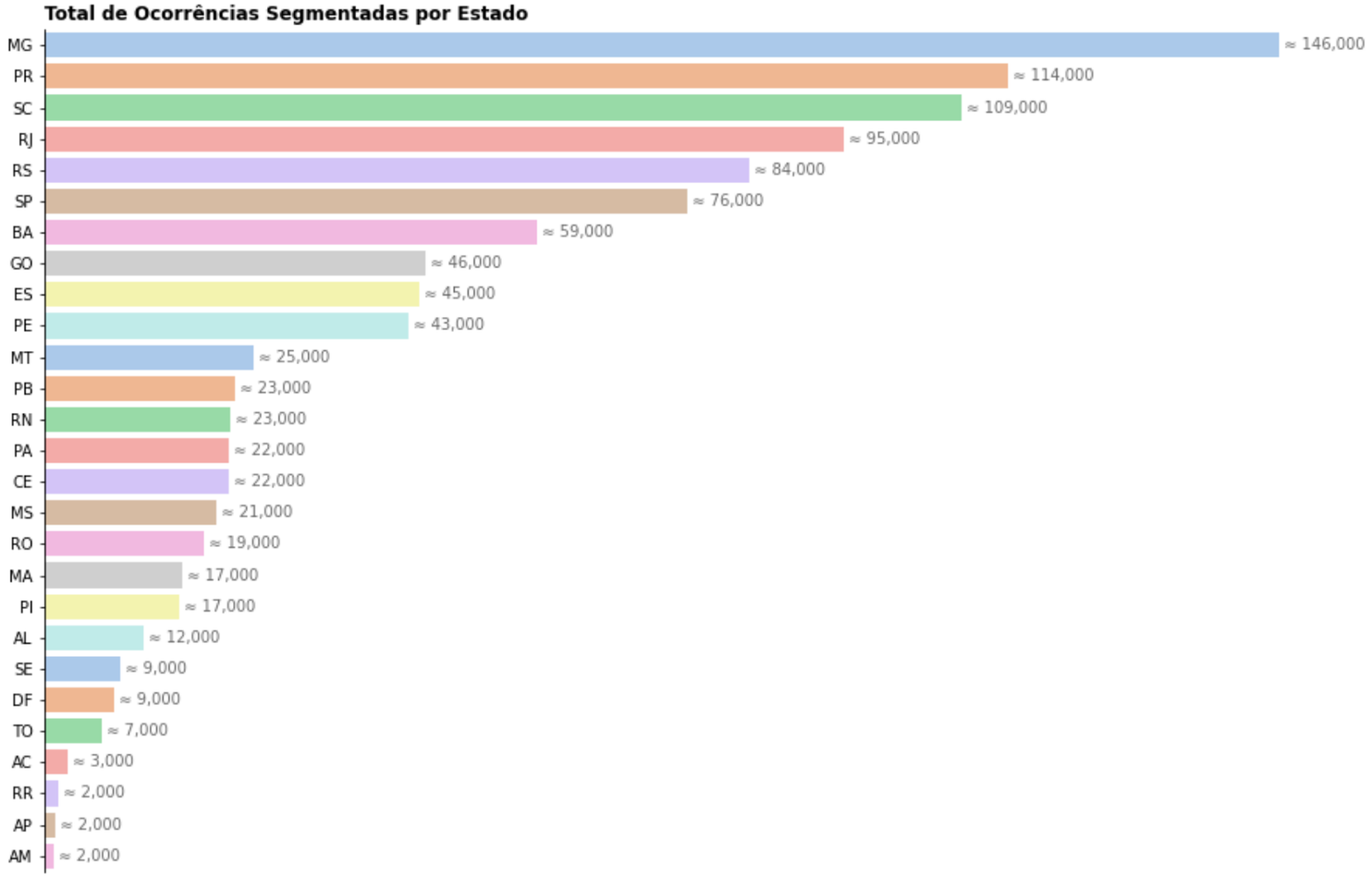
Análise Exploratória



Com os acidentes estão distribuídos em relação as horas?

INSTITUTO FEDERAL
São Paulo
Câmpus Campinas

Análise Exploratória

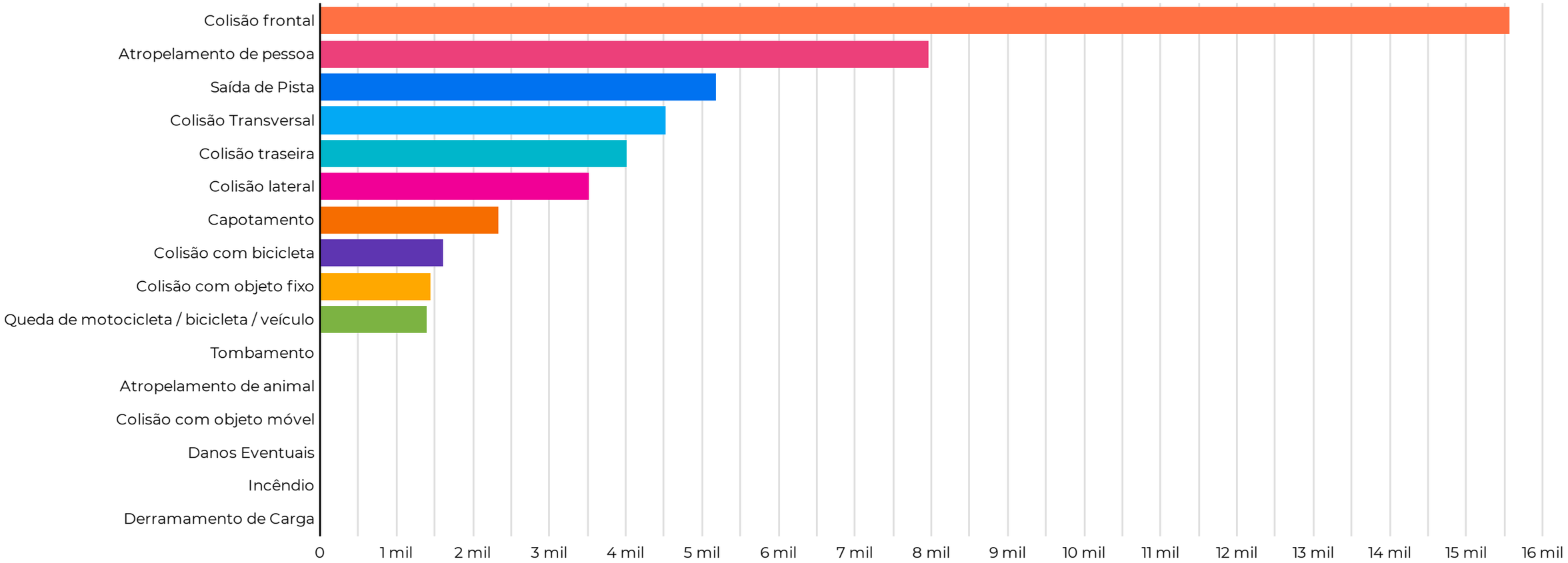


Como os ocorrências estão segmentadas por estado?

INSTITUTO FEDERAL
São Paulo
Câmpus Campinas

Análise Exploratória

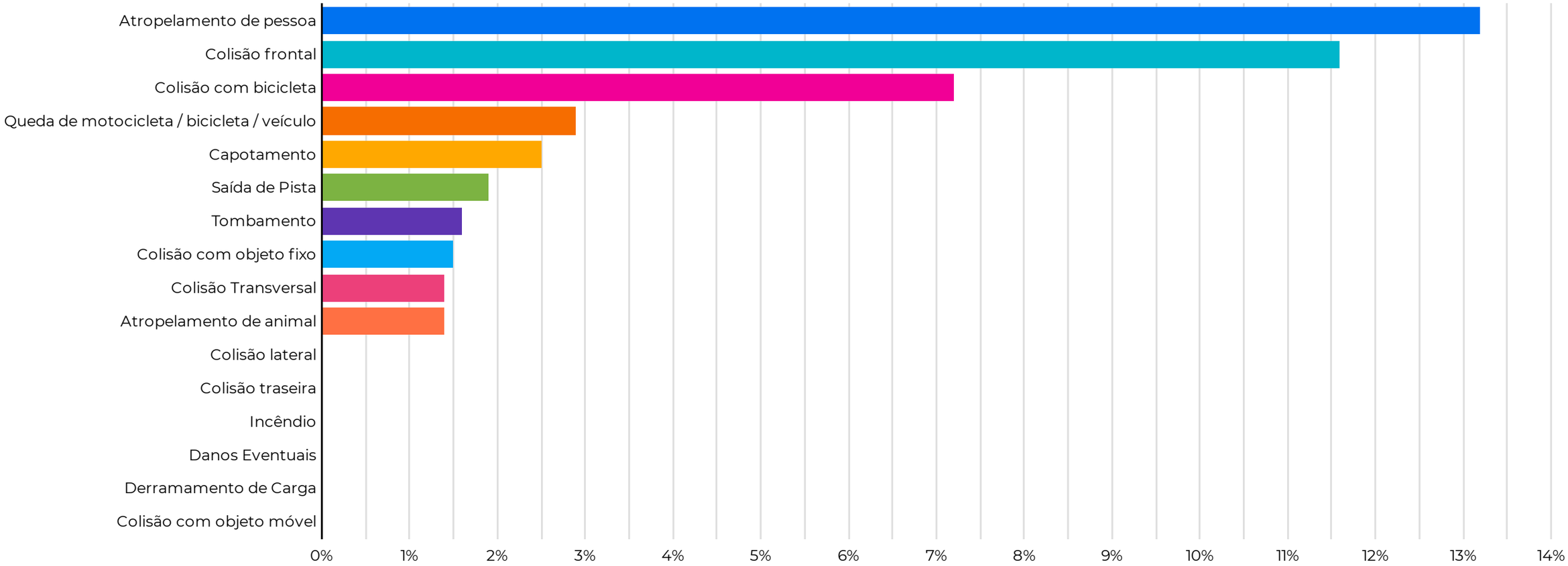
Qual tipo de acidente teve o maior número de óbitos?



INSTITUTO FEDERAL
São Paulo
Câmpus Campinas

Análise Exploratória

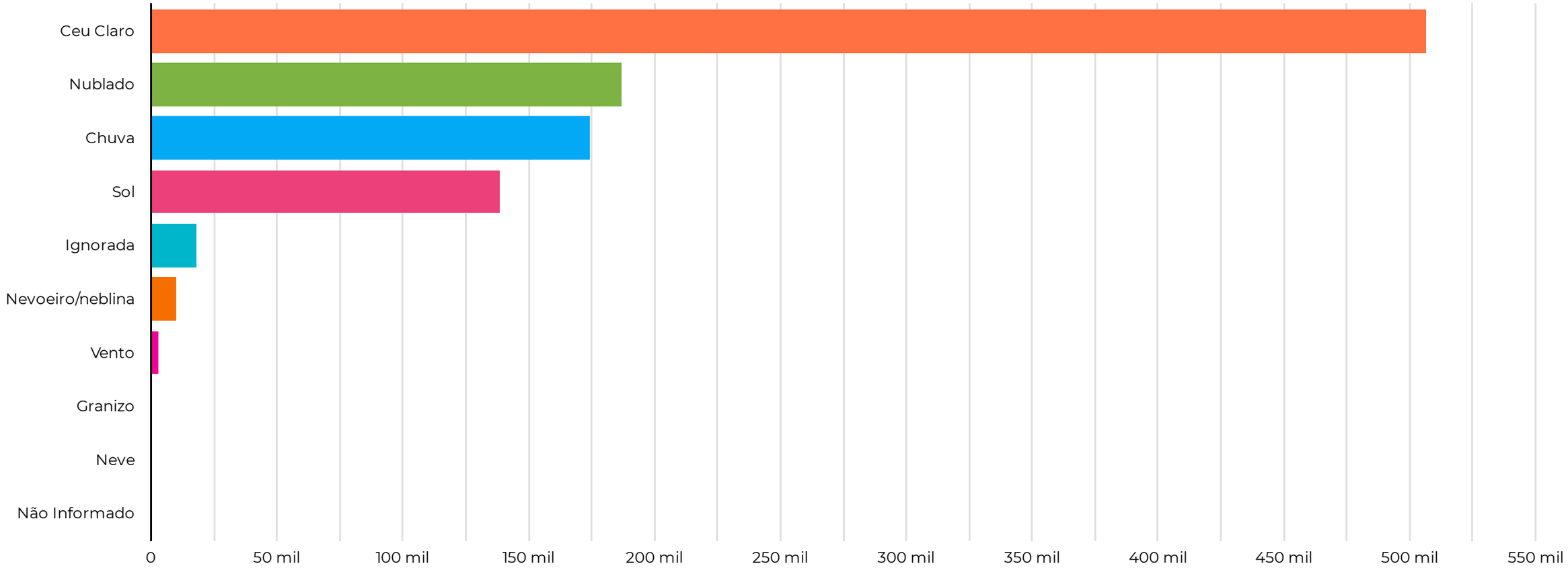
Qual a porcentagem de letalidade de cada tipo de acidente com morte?



INSTITUTO FEDERAL
São Paulo
Câmpus Campinas

Análise Exploratória

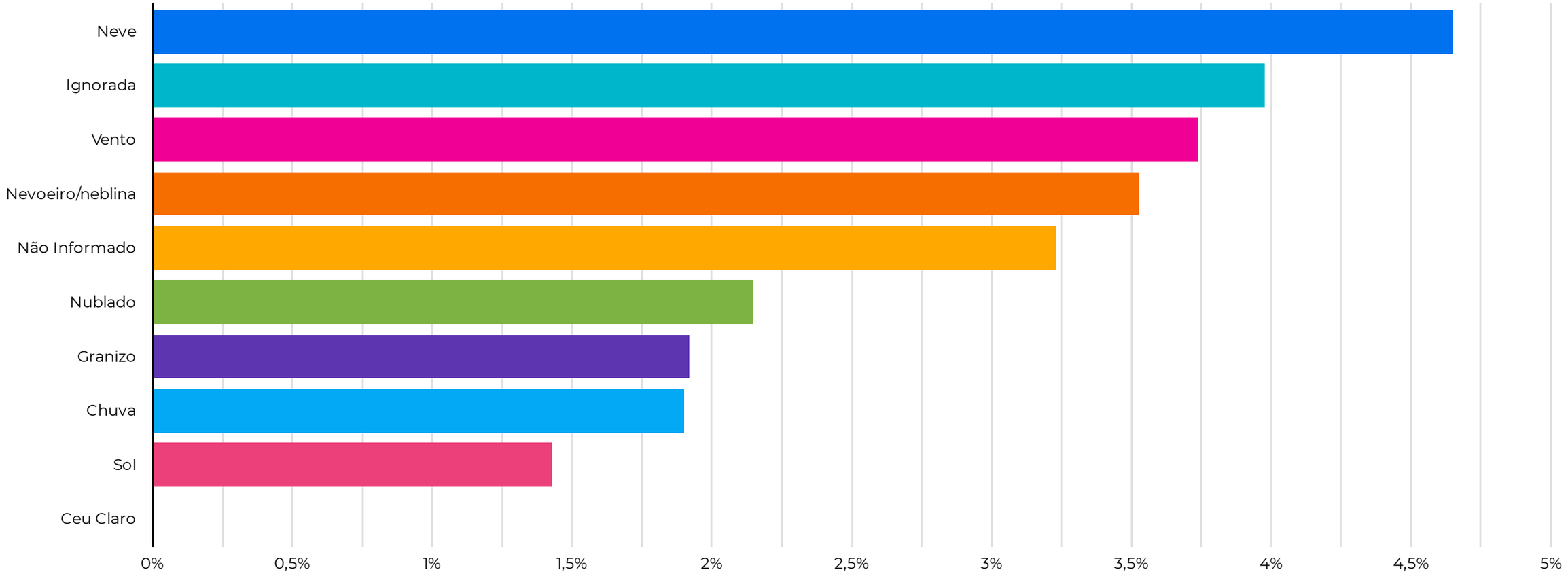
Como está distribuído a quantidade de acidentes por condição metereologica?



INSTITUTO FEDERAL
São Paulo
Câmpus Campinas

Análise Exploratória

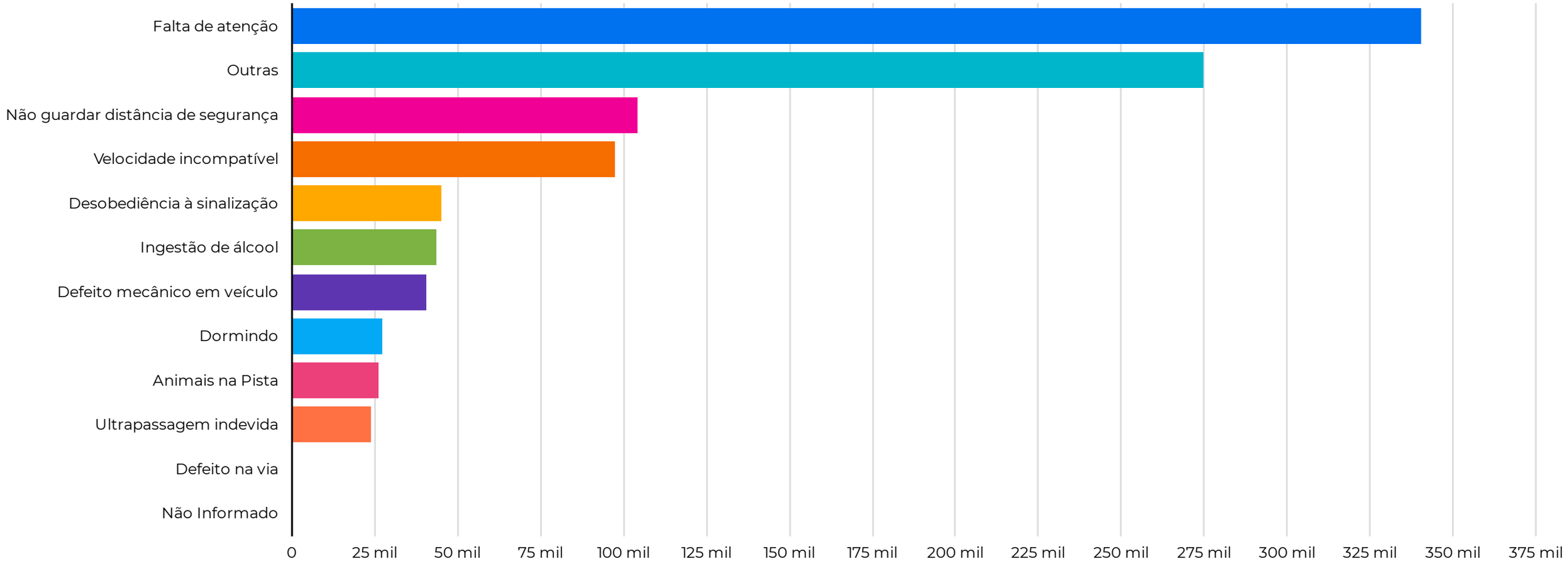
Qual a condição meteorológica mais letal?



INSTITUTO FEDERAL
São Paulo
Câmpus Campinas

Análise Exploratória

Quais são as maiores causas de acidente?



INSTITUTO FEDERAL
São Paulo
Câmpus Campinas

Classificação por Regressão Logística

- Objetivo de classificar o tipo do solo da ocorrência;
- Seleção de colunas relevantes;
- Remoção de categoria de uso do solo.

```
In [356]: df_ocorrencias_limpo.uso_solo.unique()
Out[356]: ['Rural', 'Urbano', 'Não Informado']
          Categories (3, object): ['Rural', 'Urbano', 'Não Informado']
```

```
In [357]: df_ocorrencias_limpo.groupby('uso_solo').agg({'id':'count'})
```

Out[357]:

uso_solo	id
Não Informado	10
Rural	484049
Urbano	554414

Tipos de solo encontrados na base.

INSTITUTO FEDERAL
São Paulo
Câmpus Campinas

Seleção de Variáveis Independentes

Campos	Motivo
id	Campo apenas para organização
ano	Definição temporoal, não vejo ligação direto com o problema de classificação
mes	Definição temporoal, não vejo ligação direto com o problema de classificação
data	Definição temporoal, não vejo ligação direto com o problema de classificação
hora	Definição temporoal, não vejo ligação direto com o problema de classificação
dia_semana	Definição temporoal, não vejo ligação direto com o problema de classificação
pessoas	Este campo é a soma das colunas de ilesos, feridos e mortos
feridos	Este campo é a soma das colunas feridos leves e feridos graves

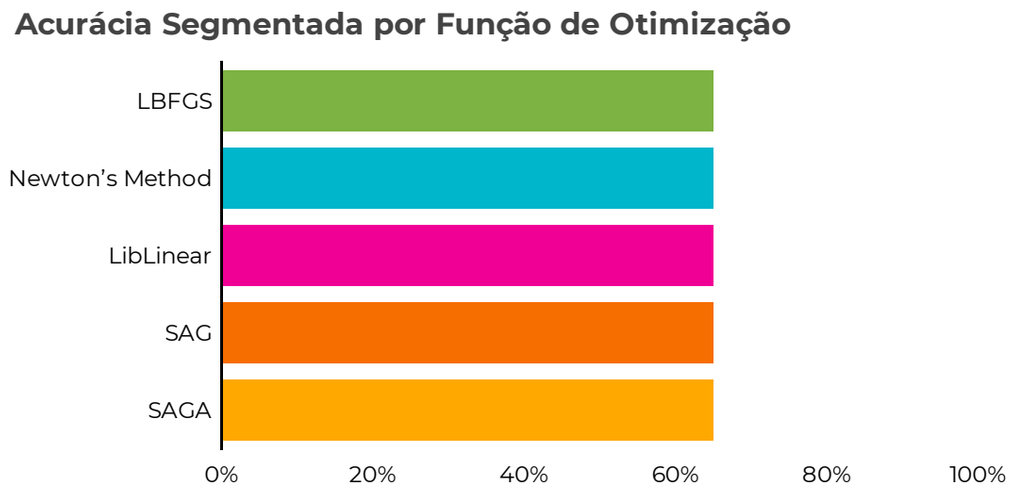
Lista de colunas retiradas da base de variáveis independentes.

INSTITUTO FEDERAL
São Paulo
Câmpus Campinas

Acurácia dos Modelos

- Testes com todas as funções de otimização disponíveis;

- Todos com o mesmo desempenho;



Conclusão

- Perfil para direcionamento de propaganda de prevenção;

- Modelo ineficiente para classificação do solo;

- Dados não são uma boa base para este problema;