PÓS-GRADUAÇÃO (*LATO SENSU*)

UNICHRISTUS

Felipe Martins Moreira - POS19100116

Lucas Nobre Bezerra - POS19100136

Raíssa Tâmara Gomes Lemine - POS19100185

Vanessa Alves Freire - POS19100186

Análise e predição do consumo de cerveja em uma região da cidade de São Paulo utilizando regressão linear

Fortaleza, Junho de 2020

1. Entendimento comercial

1.1 Definição do problema

O objetivo deste trabalho é analisar o impacto no consumo de cerveja de acordo com variações nos dias.

O conjunto de dados que foi utilizado para análise está no kaggle (https://www.kaggle.com/dongeorge/beer-consumption-sao-paulo). Ele foi coletado em São Paulo (Brasil) em uma área universitária com grupos de estudantes de 18 a 28 anos pelo período de um ano. Sete atributos foram considerados na coleta de dados.

Com base nos recursos disponibilizados, o objetivo de aprendizado de máquina é analisar os impactos das variáveis no consumo de cerveja e prever a demanda para determinados cenários.

1.2 Escopo

O escopo deste trabalho é criar um modelo de previsão de consumo com base em uma regressão linear aplicado no Python.

1.3 Plano

O TDSP (Processo de Ciência de Dados de Equipe) foi utilizado como metodologia para o presente trabalho, de forma a apresentar uma solução preditiva. O ciclo de vida descreve os principais estágios do projeto, dos quais, é descrito abaixo, a aquisição e entendimento dos dados; a modelagem, que inclui a análise exploratória, e a implantação, que apresenta um modelo de regressão linear. Nesse modelo são obtidos os coeficientes de regressão para a posterior análise gráfica do modelo preditivo.

A apresentação do código também considera os ciclos de vida do TDSP visando facilitar a compreensão do mesmo.

1.4 Equipe

O projeto foi executado por alunos do curso de pós-graduação de Ciência de Dados da Unichristus.

1.5 Métrica

O desempenho do modelo será avaliado no conjunto de testes fornecido pelos dados. A precisão é dada pelo coeficiente de relação R^2.

2. Aquisição e entendimento de dados

Para informações detalhadas, os dados estão disponíveis no Kaggle (https://www.kaggle.com/dongeorge/beer-consumption-sao-paulo).

As variáveis utilizadas são:

- Data;
- Temperatura Média (°C);
- Temperatura Mínima (°C);
- Temperatura Máxima (°C);
- Precipitação (mm);
- Final de Semana (1 = Sim; 0 = Não);
- Consumo de Cerveja (litros).

Os dados foram coletados em um período de 1 ano em São Paulo (Brasil) em uma área universitária. Foram consultados estudantes universitários com idade entre 18 e 28 anos em média. Vale ressaltar que para o teste foi considerado sempre o mesmo universo de pessoas.

3. Modelagem

Antes de iniciar o tratamento, os dados foram carregados e verificados. Foi visto que existiam algumas linhas no final da base de dados que estavam sem informação (NaN):

	Data	Temperatura Media (C)	Temperatura Minima (C)	Temperatura Maxima (C)	Precipitacao (mm)	Final de Semana	Consumo de cerveja (litros)
0	2015-01-01	27.30	23.9	32.5	0.0	0.0	25461.0
1	2015-01-02	27.02	24.5	33.5	0.0	0.0	28972.0
2	2015-01-03	24.82	22.4	29.9	0.0	1.0	30814.0
3	2015-01-04	23.98	21.5	28.6	1.2	1.0	29799.0
4	2015-01-05	23.82	21.0	28.3	0.0	0.0	28900.0
...	...	...	...	...	...	...	...
936	NaN	NaN	NaN	NaN	NaN	NaN	NaN
937	NaN	NaN	NaN	NaN	NaN	NaN	NaN
938	NaN	NaN	NaN	NaN	NaN	NaN	NaN
939	NaN	NaN	NaN	NaN	NaN	NaN	NaN
940	NaN	NaN	NaN	NaN	NaN	NaN	NaN

941 rows × 7 columns

Identificadas as linhas com este erro, foram então eliminadas.

Duas variáveis tiveram o seu tipo de dado alterado. A coluna “Data” foi alterada para formato de data e a coluna “Final de semana” foi alterada para valores inteiros, para não gerar erros na análise posterior.

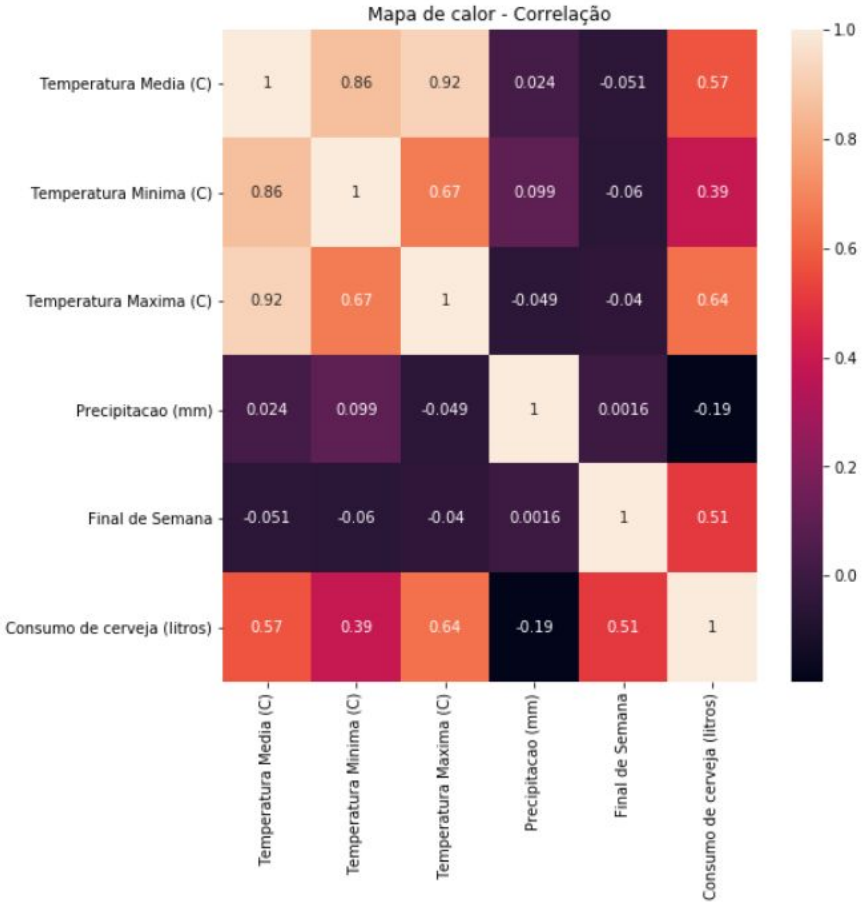
Ao final, após a limpeza, visualizamos os dados da seguinte forma:

	Data	Temperatura Media (C)	Temperatura Minima (C)	Temperatura Maxima (C)	Precipitacao (mm)	Final de Semana	Consumo de cerveja (litros)
0	2015-01-01	27.30	23.9	32.5	0.0	0	25461.0
1	2015-01-02	27.02	24.5	33.5	0.0	0	28972.0
2	2015-01-03	24.82	22.4	29.9	0.0	1	30814.0
3	2015-01-04	23.98	21.5	28.6	1.2	1	29799.0
4	2015-01-05	23.82	21.0	28.3	0.0	0	28900.0
...	...	...	...	...	...	...	...
360	2015-12-27	24.00	21.1	28.2	13.6	1	32307.0
361	2015-12-28	22.64	21.1	26.7	0.0	0	26095.0
362	2015-12-29	21.68	20.3	24.1	10.3	0	22309.0
363	2015-12-30	21.38	19.3	22.4	6.3	0	20467.0
364	2015-12-31	24.76	20.2	29.0	0.0	0	22446.0

365 rows × 7 columns

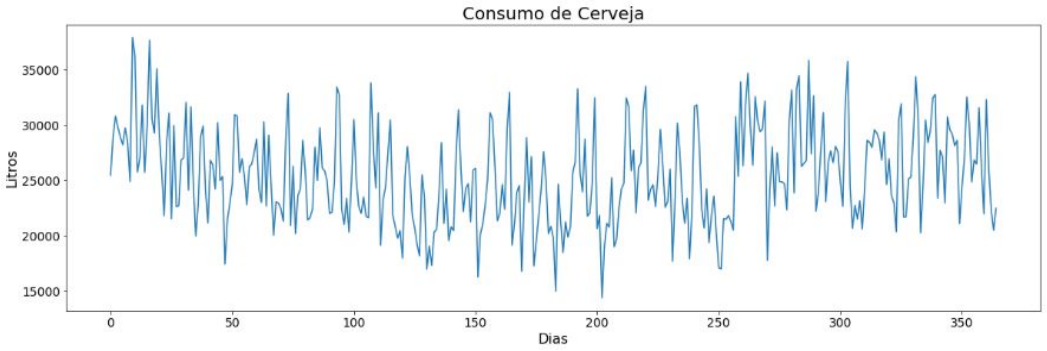
3.1 Análise exploratória

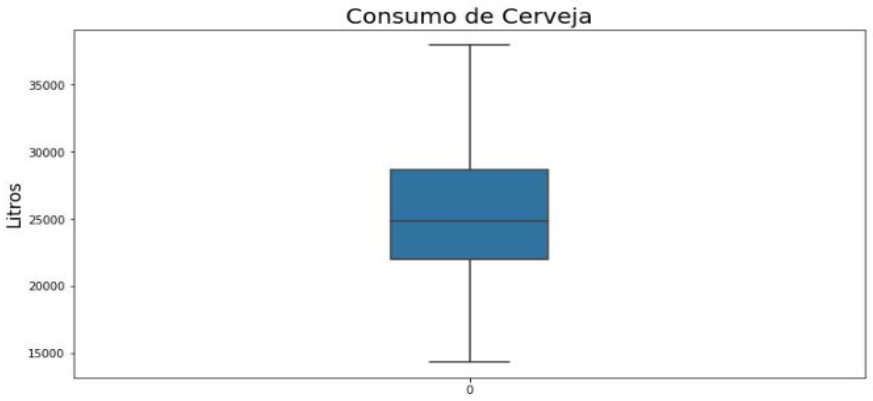
Após a limpeza dos dados, foram analisadas as estatísticas descritivas e a matriz de correlação.



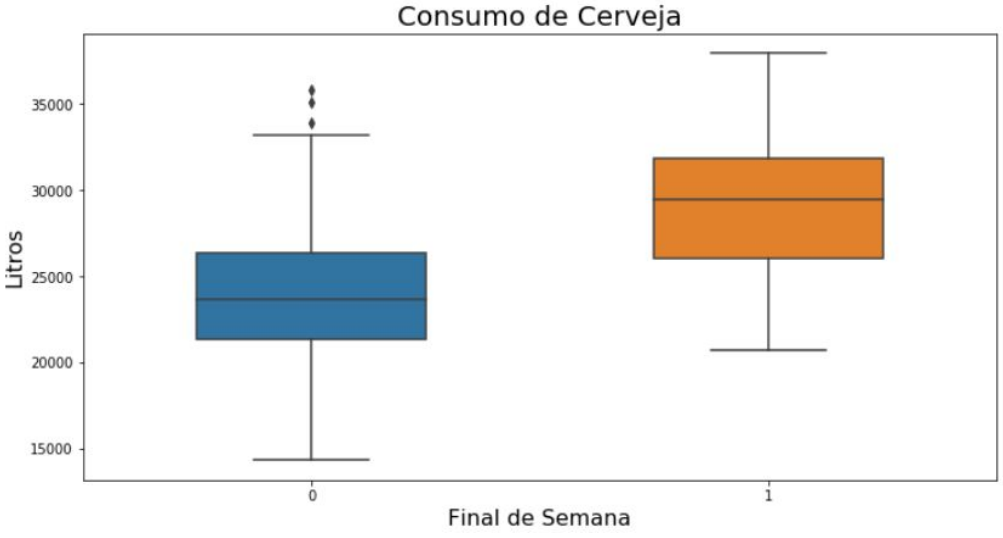
Percebemos aqui a correlação existente entre o consumo de cerveja com a variação da temperatura, precipitação e finais de semana.

A variável dependente (y) é o consumo de cerveja, pois é o que desejamos analisar de acordo com as demais variáveis. Plotada, verificamos assim o gráfico e boxplot:



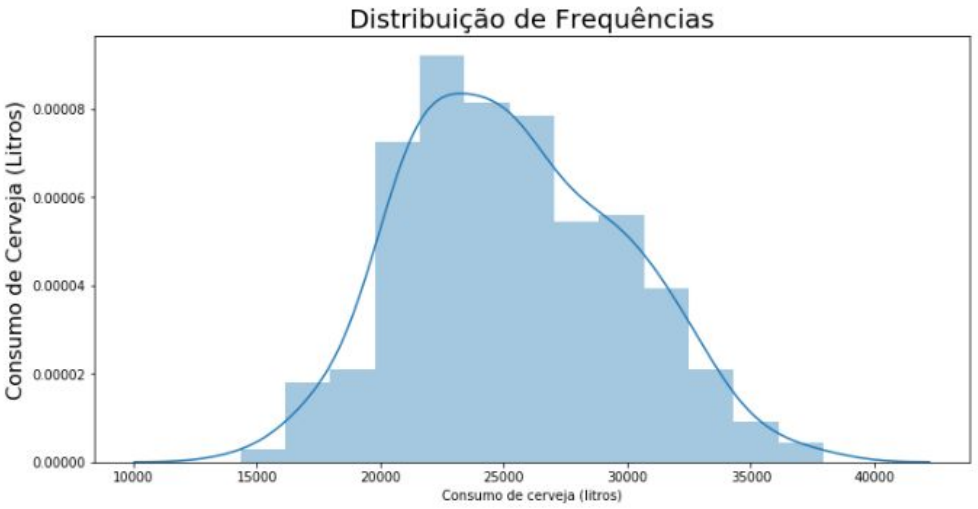


Feito isso, foi analisado o consumo por final de semana:

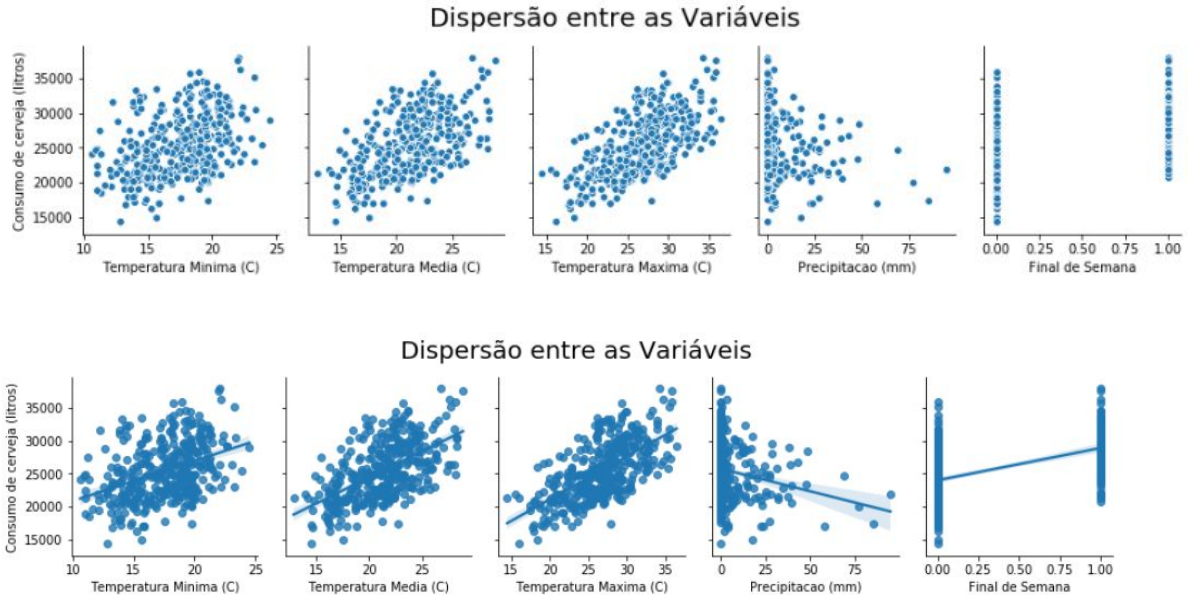


Neste gráfico, foi verificado que o consumo de cerveja é maior nos finais de semana.

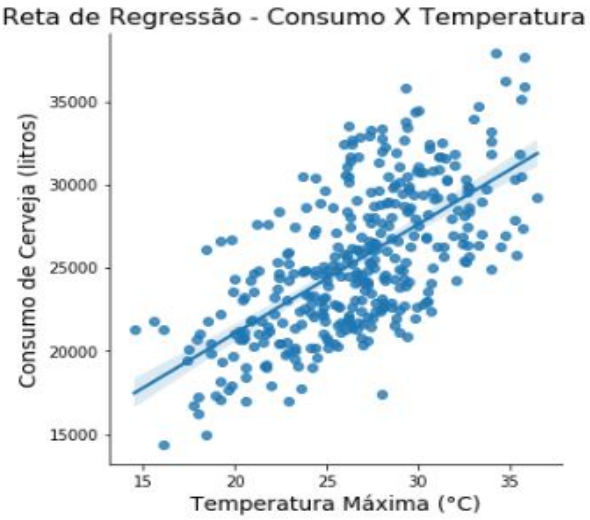
A distribuição de frequências do consumo de cerveja ficou da seguinte forma:



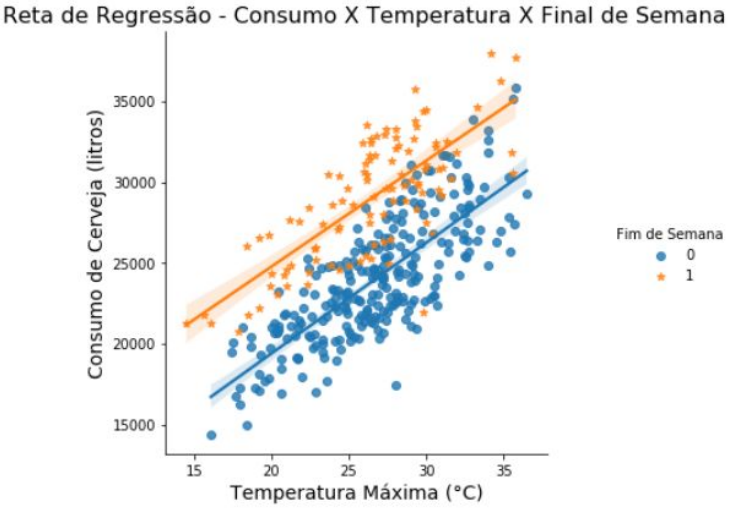
Em seguinte, plotados os gráficos de dispersão entre a Variável Dependente x Variáveis Explicativas. Fixando somente a variável de Consumo de Cerveja no eixo y, temos os gráfico de dispersão:



A reta de regressão do consumo pela temperatura máxima é:



A reta de regressão avaliando a influência da variável temperatura máxima na variável consumo separando por final de semana:



4. Desenvolvimento

4.1 Modelo Preditivo

O modelo de regressão linear foi criado seguindo os passos:

- Criando um série (pandas) para o consumo de cerveja;
- Criando um DataFrame (pandas) para armazenar as variáveis explicativas;
- Criando um dataset de treino e teste;
- Verificando o tamanho dos arquivos gerados;
- Utilizando o método fit() do objeto "modelo" para estimar nosso modelo linear utilizando os dados de treino;
- Obtendo o coeficiente de relação (R^2) do modelo estimado com os dados de TREINO;
- Gerando previsões para os dados de TESTE (X_test) utilizando o método predict() do objeto "modelo";
- Obtendo o coeficiente de determinação (R^2) para as previsões do nosso modelo.

4.2 Interpretação dos coeficientes estimados

O intercepto representa o efeito médio em Y (Consumo de Cerveja) tendo todas as variáveis explicativas excluídas do modelo. De forma mais simples, o intercepto representa o efeito médio em Y (Consumo de Cerveja) quando as variáveis Temperatura Máxima, Precipitação e Final de Semana são iguais a zero.

Significa que desconsiderando as variáveis explicativas, a média da variável dependente (Consumo) é de 5.951,97 litros.

4.3 Obtendo os coeficientes de regressão

Aplicando no modelo, verificamos que:

- Cada grau de temperatura influencia no consumo em 684,73 litros;
- Cada milímetro de precipitação varia o consumo em -60,78 litros;
- O final de semana altera o consumo em 5.401,08 litros.

Após as etapas de confirmação da ordem das variáveis explicativas no DataFrame e de criada uma lista com os nomes das variáveis do modelo, foi criado um DataFrame para armazenar os coeficientes, gerando os parâmetros:

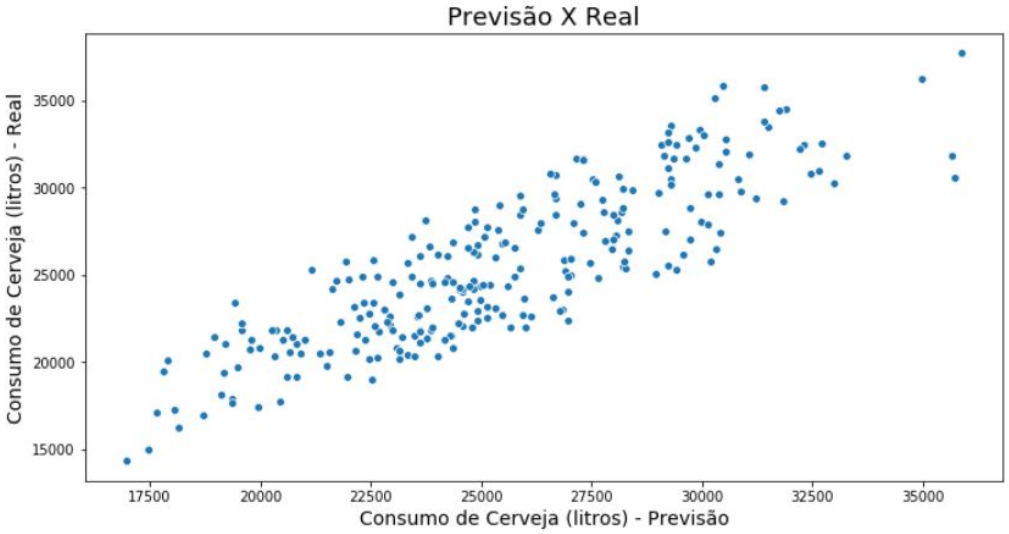
	Parâmetros
Intercepto	5951.976339
Temperatura Máxima	684.736759
Precipitação (mm)	-60.782435
Final de Semana	5401.083339

4.4 Interpretação dos coeficientes estimados

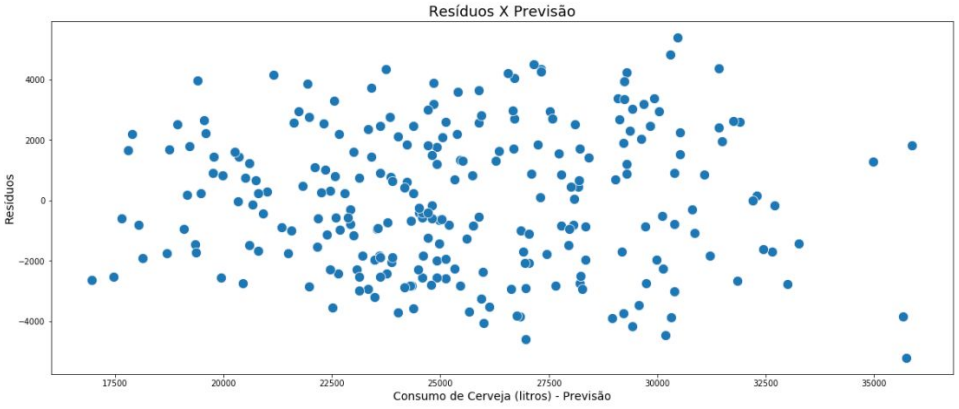
- **Intercepto** → Excluindo o efeito das variáveis explicativas (X_2=X_3=X_4=0) o efeito médio no Consumo de Cerveja seria de **5.951,98 litros**.
- **Temperatura Máxima (°C)** → Mantendo-se os valores de X_3 (Chuva) e X_4 (Final de Semana) constantes, o acréscimo de 1°C na Temperatura Máxima gera uma variação média no Consumo de Cerveja de **684,74 litros**.
- **Chuva (mm)** → Mantendo-se os valores de X_2 (Temperatura Máxima) e X_4 (Final de Semana) constantes, o acréscimo de 1mm de Chuva gera uma variação média no Consumo de Cerveja de **-60,78 litros.**
- **Final de Semana (Sim/Não)** → Mantendo-se os valores de X_2 (Temperatura Máxima) e X_3 (Chuva) constantes, o fato de o dia ser classificado como Final de Semana gera uma variação média no Consumo de Cerveja de **5.401,08 litros.**

4.5 Análises Gráficas das Previsões do Modelo

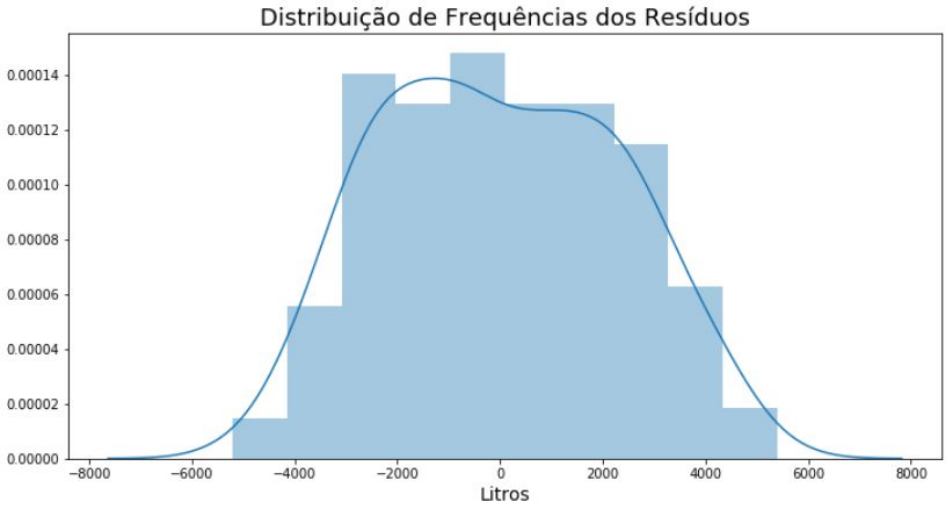
O gráfico de dispersão entre o valor estimado e o valor real é:



Os resíduos foram obtidos e plotados em um gráfico de dispersão com os valores estimados:



A distribuição da frequência de resíduos ficou então:



4.6 Salvando e testando o modelo

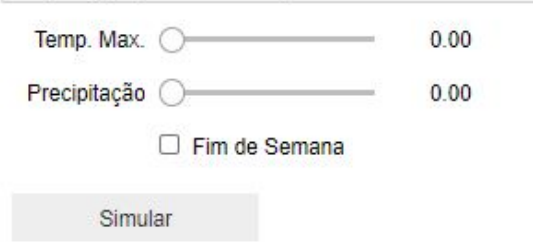
Foram introduzidos dados de entrada para gerar a previsão pontual e criar um simulador simples. O modelo estimado foi salvo para a criação de um aplicativo.

O ambiente de produção foi criado da seguinte forma:

- Foi importada a biblioteca *pickle*;
- Criados controles para temperatura, precipitação, e final de semana;
- Foram posicionados os controles;
- E foi definida a função do simulador.

Assim, pôde ser criada a interface do simulador:

```
#Chamando a interface do simulador
display(inputs, botao)
```



Foi feito um teste para exemplo. Em temperatura de 30°C, precipitação de 40mm e em final de semana, o consumo previsto na simulação é de 29463,87 litros.