Análise Inferencial de dados apresentadas pela Picmoney

Com foco em determinação de Coeficiente de correlação de Pearson (r) e Regressão linear

Após análise descritiva de todos as planilhas apresentadas pela PicMoney, realizei a análise do coeficiente de correlação de Pearson para determinar a relação entre diferentes variáveis, majoritariamente númericas, que nesse documento vai ser representado pela letra r. Após a coleta do coeficiente de correlação, optei por realizar uma análise de regressão, que consistia na comparação entre uma data variável independente (y) e uma variável dependente, a fim de realizar projeções financeiras estratégicas para a empresa, além disso serão feitas ramificações considerando outras variáveis não numéricas, a fim de obter um resultado regressivo mais palpável e direcionado. Após a análise correlativa e regressiva dos dados, disponibilizarei gráficos que representam a relação entre a equação regressiva obtida e a correlação dos dados. Dessa forma esse documento será separado da seguinte maneira:

1. Realização da análise dos dados apresentados, seguida de padronização e remapeamento de dados, caso seja necessária, de modo a possuir uma base de dados sólida, padronizada e coerente para as análises subsequentes.
2. Demonstração visual das ferramentas e padrões de código utilizados para a análise, padronização dos dados e realização dos aspectos técnicos descritos nos tópicos 3, 4 e 5.
3. Cálculo do coeficiente de correlação de Pearson (r) com os dados pré-processados na etapa anterior. Essa correlação será feita buscando inicialmente um expoente geral (Por exemplo, correlação entre o valor do cupom e o repasse feito à PicMoney) e subsequentemente uma análise mais aprofundada e subdividade (Por exemplo, correlação entre o valor do cupom e o repasse feito à PicMoney para um tipo específico de cupom). Após a determinação desses coeficientes, será realizada interpretação dos coeficientes obtidos, a fim de buscar compreender possíveis padrões e avaliação da qualidade dos dados.
4. Análise da regressão dos dados que tiveram seu Coeficiente de correlação de Pearson calculados, seguida da determinação da variável dependente (y) e da variável independente (x) para a realização do cálculo da equação da reta $(y = a + bx)$.
5. Demonstração gráfica (Diagrama de Dispersão) dos dados analisados separadamente, para compreensão visual de todos os aspectos de correlação e regressão. Por se tratar de um grande volume de dados, com a finalidade de demonstrar a dispersão dos dados serão selecionadas 15 amostras aleatórias.
6. Conclusão da análise e determinação de sua utilidade e usabilidade para demonstração e previsão de aspectos financeiros.

A padronização descrita acima será realizada de maneira sequencial para cada uma das planilhas disponibilizadas pela PicMoney. A ferramenta utilizada para a realização de todos os aspectos descritos foi a linguagem Python, por meio das bibliotecas pandas (utilizada para a leitura e acesso aos dados das planilhas), matplotlib.pyplot (utilizada para a geração dos modelos gráficos), numpy (para melhor gestão de arrays e listas de dados) e statsmodels (para o cálculo da regressão dos dados), como demonstrado no recorte a seguir:

```
#Importação da planilha que será utilizada e das bibliotecas
import pandas as pd
import numpy as np
import statsmodels.api as sm
import matplotlib.pyplot as plt
df = pd.read_csv('cupons.csv', sep = ';')
```

Além dos aspectos listados acima, também destaco que as planilhas foram renomeadas em prol de uma maior legibilidade e compreensão prática.

PicMoney-Massa_de_Teste_com_Lojas_e_Valores-10000 linhas(1) foi renomeado como compras.csv.

PicMoney-Base_Simulada_-_Pedestres_Av__Paulista-100000 linhas (1) foi renomeado como pedestres.csv.

PicMoney-Base_de_Transa__es_-_Cupons_Capturados-100000 linhas (1) foi renomeado como compras.csv.

PicMoney-Base_Cadastral_de_Players-10_000 linhas (1) foi renomeado como players.csv.

Por que a preferência por Python em relação ao R?

A escolha da linguagem Python como ferramenta principal para esta análise fundamenta-se em sua capacidade de gerenciar o ciclo de vida completo de um projeto de análise de dados, desde a padronização até a modelagem e visualização. Diferentemente do R, que é puramente estatístico, o Python é uma linguagem mais generalizado que, através de suas bibliotecas, oferece uma solução integrada, eficiente e mais versátil. Além disso, pelo fato de Python estar sendo usado em outras partes da aplicação, acredito que a opção por estar linguagem acaba oferecendo uma proposta mais eficaz, gerando prompts que podem ser utilizados em outras partes do código.

Análise dos dados presentes em cupons.csv:

A planilha cupons.csv apresenta duas variáveis fundamentalmente numéricas, são elas a coluna valor_cupom (de aqui em diante, será tratada como valor do cupom), que descreve o valor do cupom utilizado por um determinado usuário, e repasse_picmoney (de aqui em diante será tratada como repasse à PicMoney), que descreve o valor repassado para a PicMoney após a utilização do cupom. Além disso, a planilha possui variáveis não numéricas, que podem ser utilizadas para a obtenção de possíveis padrões mais detalhados, são elas tipo_cupom (que será tratada como tipo do cupom), categoria_estabelecimento (que será tratada como categoria do estabelecimento), data, hora. Além das colunas mencionadas anteriormente, a planilha também possui bairro_estabelecimento, celular, id_campanha e id_cupom, no entanto tais tabelas não serão utilizadas dentro da análise, visto que algumas dessas variáveis teriam pouca utilidade prática e poderiam gerar uma expansão considerável nos fatores analisados.

Tendo listado os fatos anteriores, tratarei de indicar algumas padronizações e remapeamentos que realizei no código, a fim de coletar dados um pouco mais generalizados. Primeiramente, optei por uma ramificação dos valores de cupons e de repasses para o PicMoney em subgrupos de acordo com o tipo de cupom. Após isso, criei variáveis que carregariam independentemente o valor total de cupons por tipo (Cashback, Produto e Desconto) e o mesmo foi feito para o repasse à PicMoney, como demonstrado no código a seguir:

```
#Ramificação de Cupons em subgrupos que representam o tipo de Cupom
soma_cashback = df[df['tipo_cupom'] == 'Cashback'][['valor_cupom', 'repasse_picmoney']].sum()
soma_produto = df[df['tipo_cupom'] == 'Produto'][['valor_cupom', 'repasse_picmoney']].sum()
soma_desconto = df[df['tipo_cupom'] == 'Desconto'][['valor_cupom', 'repasse_picmoney']].sum()

soma_cashback_valor = soma_cashback['valor_cupom']
soma_cashback_repasse = soma_cashback['repasse_picmoney']

soma_produto_valor = soma_produto['valor_cupom']
soma_produto_repasse = soma_produto['repasse_picmoney']

soma_desconto_valor = soma_desconto['valor_cupom']
soma_desconto_repasse = soma_desconto['repasse_picmoney']
```

Tendo feito, seria mais fácil de realizar correlações e regressões de maneira isolada para cada tipo de cupom utilizado.

Após isso, detectei algumas incongruências entre as colunas de categoria do estabelecimento e de nome do estabelecimento. Por exemplo, a segunda linha tratava Habib's como Lojas de Eletrônicos e a terceira linha também tratava Smart Fit como Lojas de Eletrônicos. Para contornar essa desconexão, que se repetia em outras linhas da planilha, eu optei de, por meio do Python, utilizar uma função de dicionário para renomear todos os estabelecimentos listados de acordo com sua categoria correta. Dessa forma, Habib's foi corretamente categorizado como Lanchonetes e Fast-Food, enquanto Smart Fit foi

categorizado como Academias e Studios Fitness. Essa mesma correção foi feita para todos os outros 31 estabelecimentos presentes em cupons.csv, e subsequentemente aplicado à coluna como demonstrado no código a seguir:

```
40 df['categoria_estabelecimento'] = df['nome_estabelecimento'].map(correcao_categorias).fillna(df['categoria_estabelecimento'])
```

Após isso, a última padronização necessária que foi a separação dos dados presentes em hora em períodos do dia, além da categorização de datas em dias da semana. Ou seja, horários entre 06:00 e 11:59 foram agrupados em período matutino, horários entre 12:00 e 17:59 foram agrupados em período vespertino, enquanto horários entre 18:00 e 23:59 foram agrupados em período noturno. Além disso, por meio da função datetime para classificar as datas exibidas na coluna data em dias da semana que seguiram o mapeamento ([i]:dia, ou seja [0]:domingo, [1]:segunda-feira, ..., [6]:sábado), após essa categorização foi gerada uma nova coluna no dataframe chamada dia_da_semana. Como demonstrado no código a seguir:

```
1 #Organizar Datas em dias da semana e horários em períodos do dia
2 df['data_hora'] = pd.to_datetime(df['data'] + ' ' + df['hora'], format='%d/%m/%Y %H:%M:%S')
3 df['dia_semana'] = df['data_hora'].dt.day_name()
4 df['hora_dia'] = df['data_hora'].dt.hour
5 day_names_map = {0: 'segunda-feira', 1: 'terça-feira', 2: 'quarta-feira', 3: 'quinta-feira', 4: 'sexta-feira', 5: 'sábado', 6: 'domingo'}
6 df['dia_da_semana'] = df['data_hora'].dt.dayofweek.map(day_names_map)
7 df['dia_da_semana'].value_counts().reindex(list(day_names_map.values()))
8 def classificar_periodo(hora):
9     if 6 <= hora <= 11:
10        return 'Matutino'
11    elif 12 <= hora <= 17:
12        return 'Vespertino'
13    elif 18 <= hora <= 23:
14        return 'Noturno'
15    else:
16        return 'Madrugada'
17 df['periodo_dia'] = df['hora_dia'].apply(classificar_periodo)
```

Agora, tendo os dados categorizados, corrigidos, padronizados e remapeados (Data Cleansing), Realizarei as análises de correlação e regressão distribuídas em uma análise geral (Análise direta, considerando somente repasse à PicMoney e valor do cupom), análise por tipo de cupom (Análise de regressão e correlação entre repasse à PicMoney e valor do cupom por tipo de cupom), análise por período do dia, análise por dia da semana e análise por categoria do estabelecimento.

Análise numérica e gráfica de cupons.csv:

Em Python, a correlação dos valores de cupons e seus respectivos repasses à PicMoney será feita da seguinte maneira:

```python
#Correlação entre todos o valor de todos cupons (X) e todos os repasses feitos ao PicMoney (Y)
correlacao_geral = df['repasse_picmoney'].corr(df['valor_cupom'])
print(correlacao_geral)

#Correlação entre o valor de todos os cupons cashback e repasses de cashback
correlacao_cashback = df[df['tipo_cupom'] == 'Cashback']['valor_cupom'].corr(
    df[df['tipo_cupom'] == 'Cashback']['repasse_picmoney']
)
print(correlacao_cashback)

#Correlação entre o valor de todos os cupons produto e repasses de produto
correlacao_produto = df[df['tipo_cupom'] == 'Produto']['valor_cupom'].corr(
    df[df['tipo_cupom'] == 'Produto']['repasse_picmoney']
)
print(correlacao_produto)

#Correlação entre o valor de todos os cupons desconto e repasses de desconto
correlacao_desconto = df[df['tipo_cupom'] == 'Desconto']['valor_cupom'].corr(
    df[df['tipo_cupom'] == 'Desconto']['repasse_picmoney']
)
print(correlacao_desconto)
```

Pelo código acima, foi feita a correlação entre todos os valores de cupons presentes na tabela e todos os repasses feitos à PicMoney. O mesmo padrão foi utilizado para medir a correlação entre todos os cupons de determinado tipo com seus respectivos repasses.

O cálculo do Intercepto (a) e da Inclinação (b) da Regressão Linear também foi realizado pelo Python, e teve grande ênfase na utilização da biblioteca statsmodels, que permitia a automatização do cálculo de ambos os coeficientes de regressão linear após fazer a leitura dos dados e retorno automático da Equação Geral da Reta, como demonstrado no código a seguir que retorna o valor do Intercepto e da Inclinação de todos os cupons e de cupons agrupados por tipo:

```python
#Definindo a regressão linear entre todos o valor de todos cupons (X) e todos os repasses feitos ao PicMoney (Y)
def calcular_coeficientes(grupo):
    Y = grupo['repasse_picmoney']
    X = grupo['valor_cupom']
    X = sm.add_constant(X)

    modelo = sm.OLS(Y, X).fit()

    coef_a = modelo.params['const']
    coef_b = modelo.params['valor_cupom']

    return pd.Series({'Coeficiente A': coef_a, 'Coeficiente B': coef_b})

resultados_por_grupo = df.groupby('tipo_cupom').apply(calcular_coeficientes)

resultados_geral = calcular_coeficientes(df)

print("--- Coeficientes de Regressão (A e B) Calculados Automaticamente ---\n")

print("--- Para o Geral ---")
print(f"Coeficiente A (Intercepto): {resultados_geral['Coeficiente A']:,.2f}".replace(',', 'X').replace('.', ',').replace('X', '.'))
print(f"Coeficiente B (Inclinação): {resultados_geral['Coeficiente B']:,.4f}".replace(',', 'X').replace('.', ',').replace('X', '.'))
print("-" * 40)
```

Após o cálculo tanto do Coeficiente de correlação de Pearson, quanto da Regressão, utilizei o matplotib.pyplot e sns, para a geração dos gráficos. De maneira a demonstrar de maneira gráfica e visual, a relação entre a correlação dos dados e da regressão. Como demonstrado no código a seguir, que gera o gráfico para os cupons do tipo Cashback:

```
df_cashback = df[df['tipo_cupom'] == 'Cashback']
df_amostra = df_cashback.sample(n=15, random_state=42)
sns.set_theme(style="whitegrid", palette="viridis")
Y = df_cashback['repasse_picmoney']
X = df_cashback['valor_cupom']
X = sm.add_constant(X)
modelo = sm.OLS(Y, X).fit()
r_quadrado = modelo.rsquared
x_reta = np.linspace(df_cashback['valor_cupom'].min(), df_cashback['valor_cupom'].max(), 100)
y_reta = 0.05 * x_reta
plt.figure(figsize=(10, 7))
plt.scatter(df_amostra['valor_cupom'], df_amostra['repasse_picmoney'],
            label='15 Amostras Reais de Cashback',
            s=80,
            color='#d63644',
            edgecolor='w',
            zorder=5)
plt.plot(x_reta, y_reta, color='#3b7efa',
         linewidth=2.5, label='Reta de Regressão (Y = 0.05*X)')
texto_info = f'Equação: Y = 0.05 * X\nR² = {r_quadrado:.4f}'
plt.text(0.05, 0.95, texto_info,
         transform=plt.gca().transAxes,
         fontsize=12, verticalalignment='top',
         bbox=dict(boxstyle='round,pad=0.5', fc='wheat', alpha=0.5))
plt.title('Análise de Regressão para Cupons Cashback', fontsize=18, fontweight='bold')
plt.xlabel('Valor do Cupom (R$)', fontsize=14)
plt.ylabel('Repasse PicMoney (R$)', fontsize=14)
plt.legend(fontsize=12)
plt.tight_layout()
plt.show()
```

Esse será o padrão utilizada para o cálculo do Coeficiente de correlação de Pearson e da Regressão durante todo o trabalho, então a partir de agora irei me abster da demonstração da codificação utilizada para focar, de fato, na análise numérica, gráfica e descritiva dos dados estudados e avaliados.

Comparação entre todos os valores de cupons e repasses à PicMoney:

Para realizar todas as análises da planilha cupons.csv, utilizarei o valor dos cupons como variável independente (x) e o repasse à PicMoney como variável dependente (y), dessa forma todas as análises preditivas realizadas utilizarão o valor dos cupons para prever o possível repasse que será feita à PicMoney.

Análise de valor cupom x repasses

$r = 0.36899$ ...

Como r → 0,369, utilizaremos 0,369 para descrever e representar o coeficiente de correlação.

Com esse valor, também poderemos também inferir o valor de R^2, que será calculado pelo valor de r elevado ao quadrado, dessa forma obtemos que:

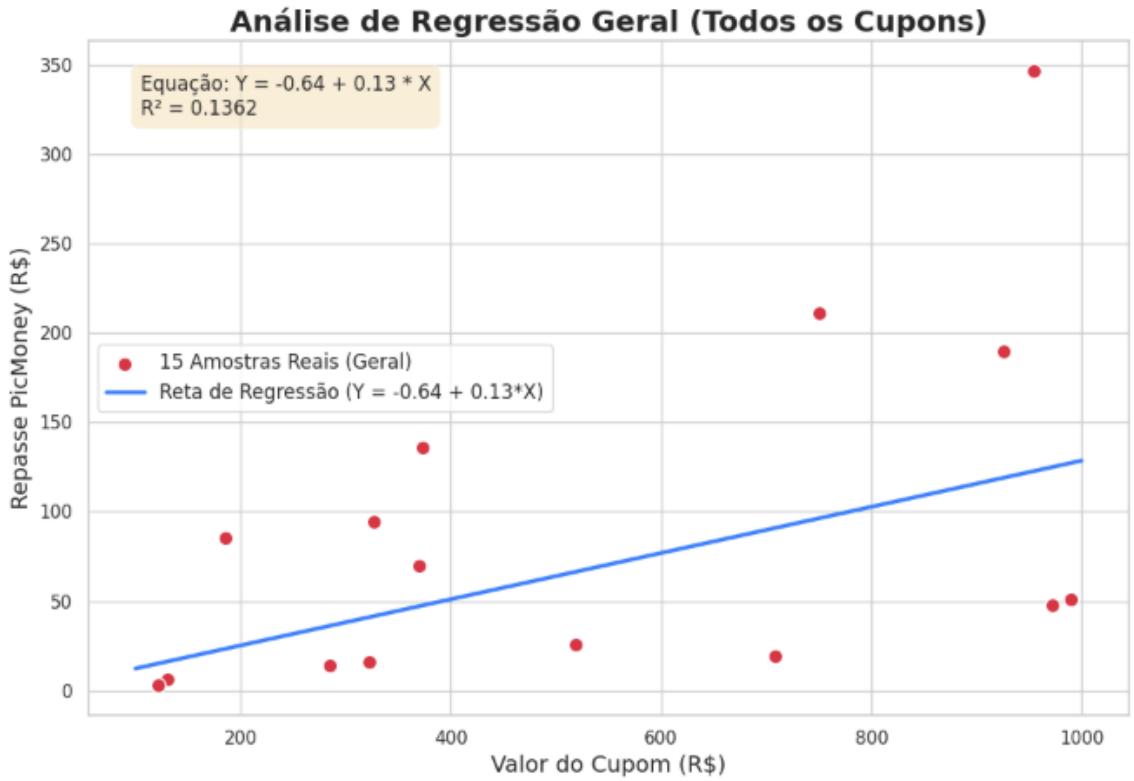
$$\boldsymbol{R^2 = (0,369)^2 = 0,1362}$$

Agora para finalizar a análise numérica, vamos calcular a regressão desse conjunto de dados, de modo a obter que:

$$\boldsymbol{y = -0,64 + 0,13x}$$

Os resultados obtidos anteriormente indicam a existência de uma correlação linear positiva, porém fraca, entre as duas variáveis, com um coeficiente (r) de 0,369. Essa noção é corroborada também pelo coeficiente de determinação (R^2) de 0,1362, demonstrando que apenas 13,62% da variabilidade no repasse à PicMoney por cupom utilizado pode ser atrelada à variação do valor do cupom.

A equação de regressão obtida, demonstrada $y = -0,64 + 0,13x$ sugere que para cada R$ acrescido ao valor do cupom, há um acréscimo de 0,13R$ no repasse para a PicMoney, começando a partir de -0,64R$. No entanto, apesar de ter um certo valor preditivo, por conta da baixa correlação, muitos dos valores testados empiricamente pela equação acima vão acabar estando em dissonância do valor real, como será demonstrada pelo gráfico a seguir:



Dessa forma, de fato, o poder preditivo do repasse à PicMoney para valores de cupons generalizados acaba sendo bem limitado, por conta da baixa

correlação direta entre essas variáveis. A fim de obter resultados mais legítimos, recomenda-se a consideração de outras variáveis.

Análise de valor cupom x repasses por tipo de cupom

Como a análise feita anteriormente não indica índices muito práticos e preditivos, realizaremos a mesma análise feita anteriormente, mas ramificada por tipos diferentes de cupons, com a finalidade de obter variáveis com um ajuste de maior qualidade, e, portanto, de maior potencial avaliativo. Analisarei, respectivamente, Cashback, Produto e Desconto.

Análise de valor cupom x repasses (Cashback)

$\boldsymbol{r = 0,9999}$ ...

Como, $r \rightarrow 1$, utilizaremos 1 para descrever e representar o coeficiente de correlação.

Com esse valor, também poderemos também inferir o valor de R^2, que será calculado pelo valor de r elevado ao quadrado, dessa forma obtemos que:

$\boldsymbol{R^2 = 1^2 = 1}$

Agora para finalizar a análise numérica, vamos calcular a regressão desse conjunto de dados, de modo a obter que:

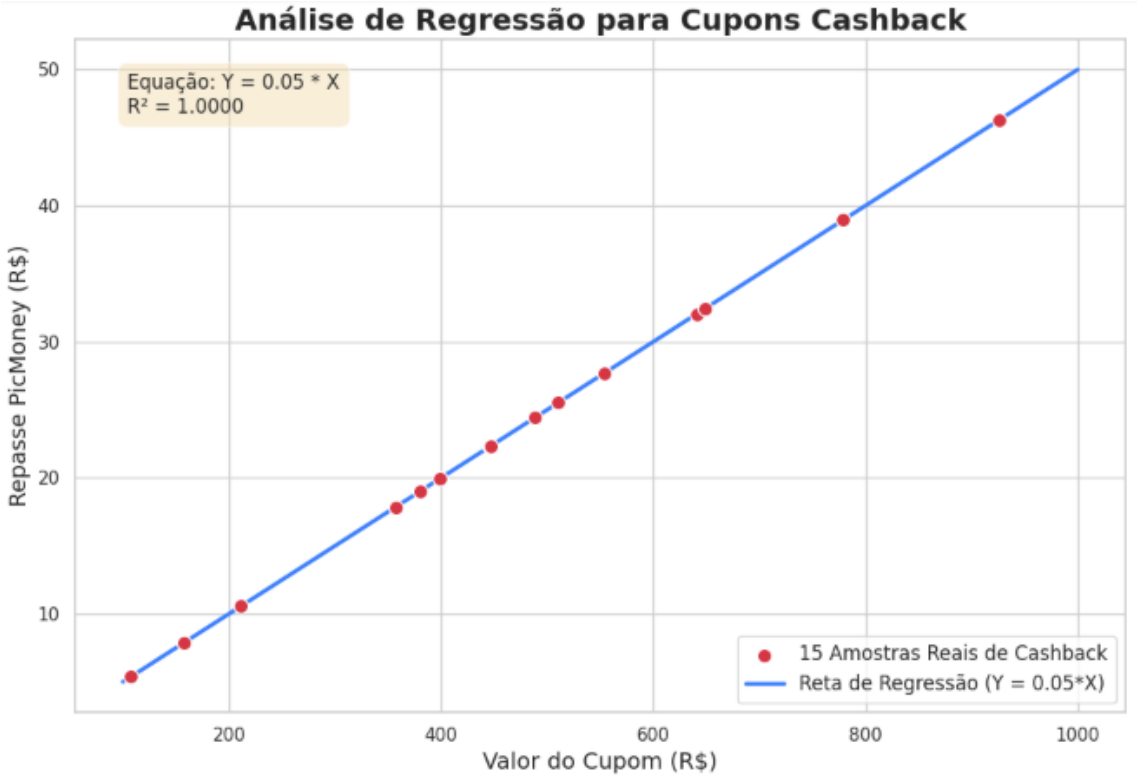
$\boldsymbol{y = 0,05x}$

Os resultados obtidos indicam uma correlação linear positiva, praticamente perfeita entre as duas variáveis, com um coeficiente (r) que tende consideravelmente a 1, que foi o valor utilizado para a análise. Tal concepção é também corroborada pelo coeficiente de determinação (R^2), demonstrando que 100% da variabilidade no repasse à PicMoney por cupom de Cashback utilizado pode ser atrelada à variação do valor do cupom.

A equação de regressão obtida, demonstrada $y = 0,05x$ sugere que para cada R$ acrescido ao valor do cupom, há um acréscimo de 0,05R$ no repasse para a PicMoney. Devido à correlação de um para a um, todos os valores testados empiricamente pela equação acima vão acabar estando em conluio com o valor real, como será demonstrado pelo exemplo e pelo gráfico a seguir:

Linha dois da tabela cupons.csv (id_cupom = CUP542835) tem um cupom cashback de valor R$ 229.64

$y = 229,64 * 0,05 = 11.48$, o mesmo repasse à PicMoney presente na linha dois.



Como demonstrado nos exemplos à cima, há um poder preditivo elevadíssimo para a análise de cupons cashback. Se tratando de um modelo que pode, sem maiores problemas, ser utilizado na estratégia financeira da empresa para determinar resultados ou metas financeiras de maneira categórica.

Análise de valor cupom x repasses (Produto)

$$\boldsymbol{r = 0,7225}$$

Como, $r \rightarrow 0{,}7225$, utilizaremos 0,7225 para descrever e representar o coeficiente de correlação.

Com esse valor, também poderemos também inferir o valor de R^2, que será calculado pelo valor de r elevado ao quadrado, dessa forma obtemos que:

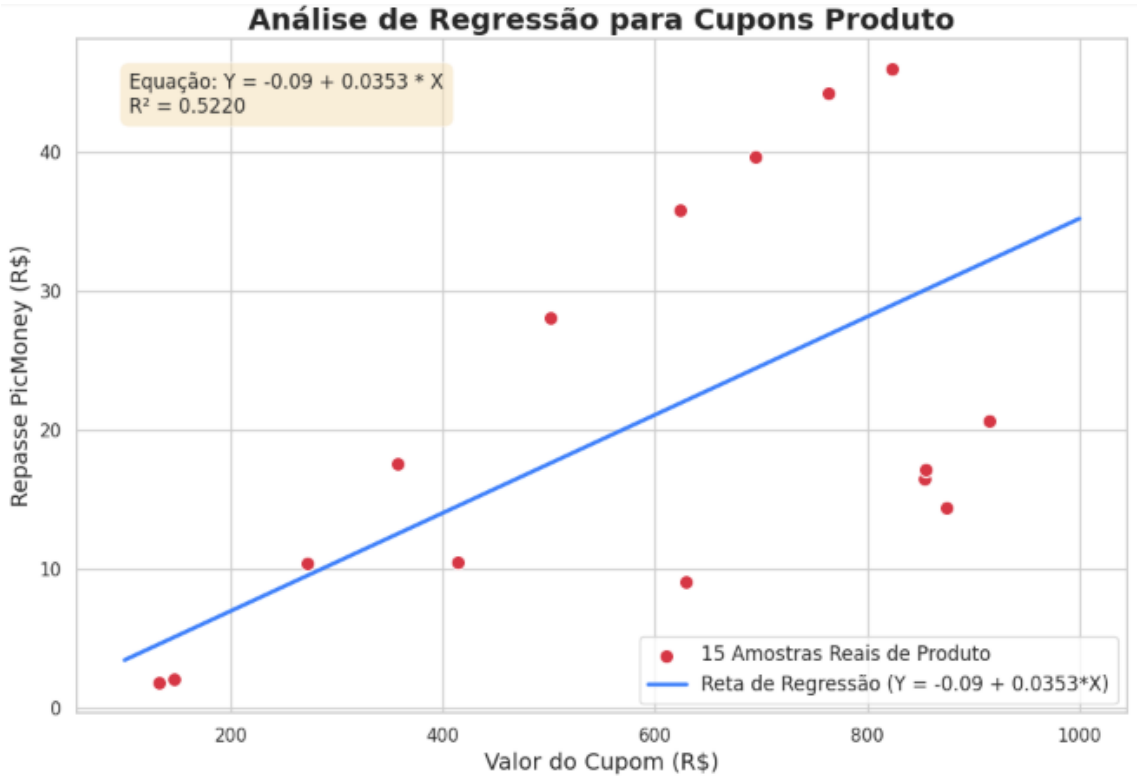
$$\boldsymbol{R^2 = 0,7225^2 = 0,522}$$

Agora para finalizar a análise numérica, vamos calcular a regressão desse conjunto de dados, de modo a obter que:

$$\boldsymbol{y = -0,09 + 0,0353x}$$

Ao segmentar a análise para os cupons do tipo Produto, os resultados obtidos indicam a existência de uma correlação linear positiva forte entre as duas variáveis, com um coeficiente (r) de 0,7225. Essa noção é corroborada também pelo coeficiente de determinação (R2) de 0,522, demonstrando que 52,2% da variabilidade no repasse à PicMoney pode ser atrelada à variação do valor do cupom neste segmento específico.

A equação de regressão obtida, demonstrada $y = -0{,}09 + 0{,}0353x$ sugere que para cada R$ acrescido ao valor do cupom, há um acréscimo de 0,0353R$ no repasse para a PicMoney, começando a partir de -0,09R$. Ao contrário da análise generalizada, que possuía uma correlação fraca, o valor preditivo do modelo para cupons de produto é consideravelmente mais robusto. Por conta da forte correlação, os valores previstos pela equação de regressão tenderiam a apresentar maior consonância com os valores reais, tornando o valor do cupom um indicador mais confiável e útil para prever o valor do repasse, como demonstrado pelo gráfico a seguir:



Como demonstrado no gráfico, há um poder preditivo consideravelmente maior para a análise de cupons de produto, em comparação com o modelo generalizado. A menor dispersão dos pontos de dados em torno da linha de regressão evidencia uma relação muito mais consistente e previsível, apesar de não ser tão precisa quanto o modelo de Cashback, como pode ser visto pelo fato de algumas amostras estarem muito próximas da reta regressiva, enquanto algumas estão um pouco mais distantes.

Portanto, trata-se de um modelo com uma boa relevância estratégica, que pode ser utilizado para aprimorar o planejamento financeiro da empresa, no entanto poderia ser valido considerar outras variáveis, para se obter uma precisão tão boa quanto a dos cupons de cashback. Com base nesta análise, é possível estimar com maior confiança os repasses que serão gerados a partir de

campanhas de produto e, consequentemente, estabelecer metas financeiras mais precisas e realistas para este segmento específico.

Análise de valor x repasses (Desconto)

$\boldsymbol{r = 0,7426}$ ...

Como, r → $0{,}7426$, utilizaremos 0,7426 para descrever e representar o coeficiente de correlação.

Com esse valor, também poderemos também inferir o valor de R^2, que será calculado pelo valor de r elevado ao quadrado, dessa forma obtemos que:

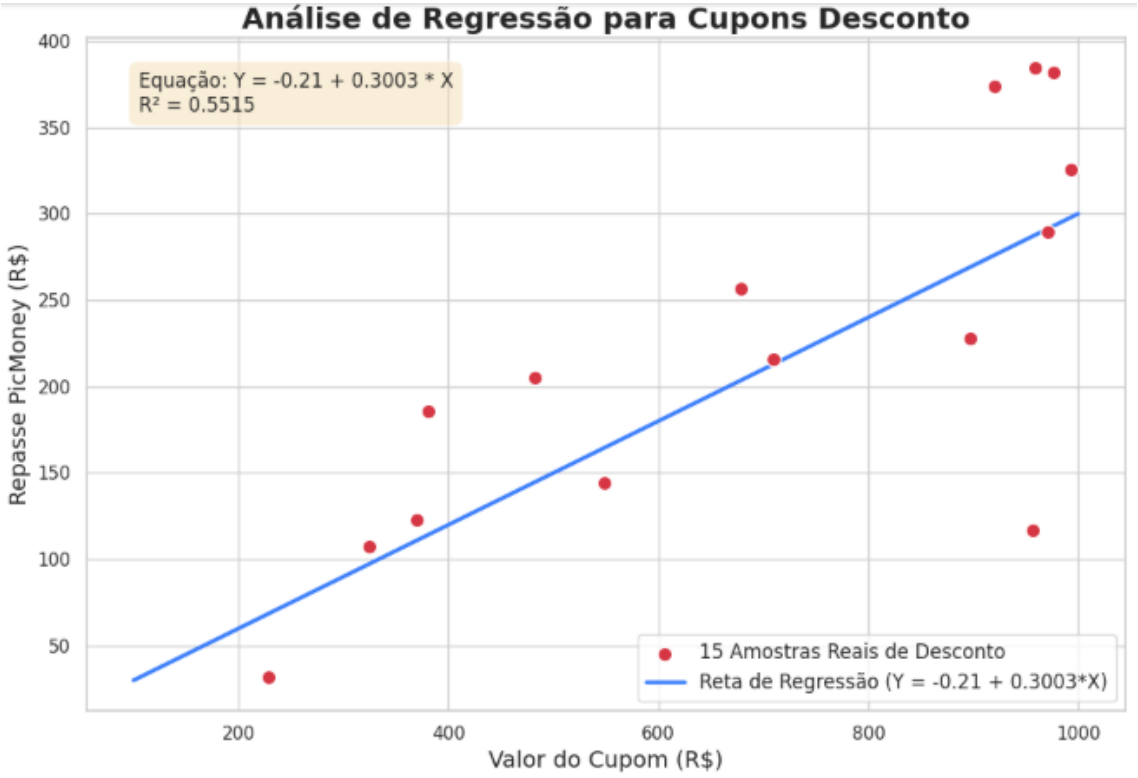
$\boldsymbol{R^2 = (0,7426^2) = 0,551}$

Agora para finalizar a análise numérica, vamos calcular a regressão desse conjunto de dados, de modo a obter que:

$\boldsymbol{y = -0,21 + 0,3003x}$

Ao subdividir a análise para os cupons do tipo Desconto, os resultados obtidos indicam a existência de uma correlação linear positiva forte entre as duas variáveis, com um coeficiente (r) de 0,7426, um pouco mais forte do que a correlação dos cupons de tipo Produto. Essa noção é corroborada também pelo coeficiente de determinação (R^2) de 0,5515, demonstrando que 55,15% da variabilidade no repasse à PicMoney pode ser atrelada à variação do valor do cupom neste segmento específico.

A equação de regressão obtida, demonstrada $y = -0{,}21 + 0{,}3003x$ sugere que para cada R$ acrescido ao valor do cupom, há um acréscimo de 0,3003R$ no repasse para a PicMoney, começando a partir de -0,21R$. Ao contrário da análise generalizada, que possuía uma correlação fraca, o valor preditivo do modelo para cupons de desconto é consideravelmente mais robusto. Por conta da forte correlação, os valores previstos pela equação de regressão tenderiam a apresentar maior consonância com os valores reais, tornando o valor do cupom um indicador mais confiável e útil para prever o valor do repasse, como demonstrado pelo gráfico a seguir:



Tendo em vista o gráfico de desconto, há um poder preditivo bastante semelhante em comparação com o modelo de cupons de produto.

Portanto, trata-se de um modelo que, assim como o modelo de produto, apresenta uma boa relevância estratégica e que pode ser utilizado para visualizar o planejamento financeiro da empresa, no entanto poderia ser valido considerar outras variáveis, para se obter uma precisão tão boa quanto a dos cupons de cashback. Também é nítido que a regressão para os cupons de desconto possui uma inclinação consideravelmente maior, quase 10 vezes maior, do que a inclinação dos cupons de produto.

Análise de valor cupom x repasse por período do dia:

A seguir, realizarei a análise do coeficiente de correlação, do coeficiente de determinação e da regressão considerando tipos os três períodos do dia que eu havia definido anteriormente e também no meu código em Python, são eles período matutino (06:00-11:59), período vespertino (12:00-17:59) e período noturno (18:00-23:59). O período da madrugada foi desconsiderado, pois a planilha não dispõe de movimentações nos horários entre 00:00 e 05:59.

Esta análise busca compreender se, assim como o tipo de cupom, o período do dia em que um cupom é utilizado acaba impactando no repasse que será feito à PicMoney, possibilitando uma maior capacidade de análise inferencial e preditiva dos dados analisados. Demonstrarei os dados por meio de uma tabela, pois ao se realizar a análise, percebe-se um padrão muito semelhante ao padrão presente na análise geral realizada no início.

	Manhã	Tarde	Noite
Coeficiente de correlação (r)	0,370	0,369	0,366
Coeficiente de determinação (R^2)	0,1369	0,1362	0,1342
Regressão Linear	$y = 0,64 + 0,1304x$	$y = -0,85 + 0,1291x$	$Y= -0,36 + 0,1277x$

Conclui-se, portanto, que, por todos os índices obtidos, é nítido que tanto o coeficiente de correlação, quanto o coeficiente de determinação, quanto a regressão linear seguem um padrão muito semelhante. Dessa forma, é possível inferir que o período do dia tem uma influência muito pequena nos valores de repasse feitos à PicMoney, reforçados por um baixo índice de correlação e um índice de determinação ainda menor, esses fatores tornam muito difícil e impreciso qualquer possibilidade de predição.

Análise de valor cupom x repasse por dia da semana:

	Coeficiente de correlação (r)	Coeficiente de determinação (R^2)	Regressão Linear
Domingo	0,3639	0,1324	$y = -0,27 + 0,1227x$
Segunda	0,3629	0,1317	$y = 0,66 + 0,1274x$
Terça	0,3662	0,1341	$y = 0,33 + 0,1287x$
Quarta	0,3673	0,1349	$y = -0,89 + 0,1294x$
Quinta	0,3760	0,1340	$y = -1,46 + 0,1319x$
Sexta	0,3704	0,1372	$y = -0,84 + 0,1284x$
Sábado	0,3789	0,1430	$y = -2,52 + 0,1338x$

Ao analisar os índices de coeficiente de correlação, coeficiente de determinação e de regressão linear, considerando os diferentes dias da semana, também há uma clara aproximação e semelhança com os dados obtidos na tabela anterior, o que reforça que se trata de outra variável muito pouco influente na determinação do repasse feito à PicMoney.

Análise de valor cupom x repasse por categoria de estabelecimento:

	Coeficiente de correlação (r)	Coeficiente de determinação (R^2)	Regressão Linear
Academias e Studios Fitness	0,3661	0,1340	$y = 0{,}20 + 0{,}1271x$
Cafeterias e Bistrôs Modernos	0,3723	0,1386	$y = -1{,}20 + 0{,}1314x$
Clubes e Centros de Convivência	0,3754	0,1409	$y = -1{,}84 + 0{,}1355x$
Clínicas Médicas e Laboratórios	0,3705	0,1372	$y = -1{,}36 + 0{,}1335x$
Espaços Culturais e de Experiência Interativa	0,3828	0,1465	$y = -5{,}72 + 0{,}1401x$
Farmácias e Drogaria	0,3642	0,1326	$y = 0{,}09 + 0{,}1268x$
Lanchonetes e Fast-Food	0,3679	0,1353	$y = -0{,}69 + 0{,}1271x$
Lojas de Eletrodomésticos e Utilidades Domésticas	0,3601	0,1296	$y = 0{,}60 + 0{,}1237x$
Lojas de Roupas e Calçados	0,3766	0,1418	$y = -1{,}85 + 0{,}1319x$
Restaurantes e Gastronomia	0,3699	0,1368	$y = 0{,}34 + 0{,}1282x$
Supermercados de Bairro e Mercadinhos	0,3688	0,1306	$y = -1{,}34 + 0{,}1301x$
Supermercados e Mercados Express	0,3595	0,1292	$y = 2{,}86 + 0{,}1223x$

Por fim, ao comparar os valores de repasse gerados por valor de cupom para determinada categoria de estabelecimento, também se torna nítido que se trata de outra variável pouco relevante na predição do percentual de repasse recebido pela PicMoney, e que também varia muito pouco em relação aos índices obtidos nas tabelas anteriores e na primeira análise feita considerando somente o repasse (y) e o valor do cupom (x).

Ao analisar as tabelas anteriores, considerei ramificá-las em subgrupos maiores, tais quais valor x cupom por tipo de cupom e período do dia, valor x cupm por tipo de cupom e categoria do estabelecimento, mas não creio que seja válido e razoável, pois a tendência, por conta da pouca relevância dos índices obtidos anteriormente, é que os resultados sejam muito semelhantes aos obtidos nas comparações considerando somente o dia de cupom e nas comparações feitas nas três tabelas anteriores. Portanto, exigiriam um desgaste maior, estenderiam consideravelmente o tamanho do documento e resultariam em um resultado pouco satisfatório.

Conclusão sobre cupons.csv:

A análise inferencial realizada sobre a base de dados cupons.csv teve como objetivo compreender a relação entre o valor dos cupons utilizados pelos

usuários e o respectivo repasse financeiro à PicMoney, a fim de criar modelos preditivos para a estratégia da empresa.

Os dados avaliados demonstraram que uma análise generalizada, considerando o volume total de cupons, é inadequada e leva a conclusões equivocadas. O modelo geral apresentou uma correlação linear positiva muito fraca (r = 0,369) e um poder de explicação baixíssimo (R^2 = 0,1362), tornando-o um instrumento impreciso e de pouca praticidade para previsões financeiras.

Uma parte fundamental da análise foi a tentativa de refinar o modelo por meio da segmentação por outras variáveis. No entanto, as comparações por período do dia, dia da semana e categoria do estabelecimento mostraram-se pouco eficazes. Conforme detalhado nas tabelas de análise, os coeficientes de correlação e determinação para todos esses subgrupos permaneceram consistentemente baixos, com valores quase idênticos aos do fraco modelo geral. Isso comprova que tais variáveis possuem uma influência estatística ínfima na determinação do repasse e não agregam qualquer valor preditivo.

A principal descoberta desta análise jaz na segmentação pela variável tipo_cupom, que se revelou o único fator verdadeiramente determinante para o comportamento do repasse. Podendo se concluir que:

1. Cupons Cashback: A relação é praticamente perfeita ($r \approx 1,0$; $R^2 = 1,0$), indicando um modelo determinístico onde o repasse corresponde a uma taxa fixa de 5% sobre o valor do cupom. Este modelo é extremamente confiável para projeções financeiras.
2. Cupons Produto e Desconto: Ambos os tipos apresentaram uma correlação linear positiva forte (r = 0,7225 e r = 0,7426, respectivamente), com seus modelos explicando mais de 50% da variabilidade do repasse (R^2 = 0,522 e R^2 = 0,551). Isso os torna modelos preditivos consideravelmente robustos e estrategicamente relevantes, com destaque para a taxa de repasse dos cupons de desconto (y = -0,21 + 0,3003x), que se mostrou quase dez vezes superior à dos cupons de Produto.

Portanto, conclui-se que a estratégia financeira e de previsão de receita da PicMoney deve descartar abordagens generalistas ou baseadas em fatores de baixo impacto como horário ou local. O foco deve ser em uma abordagem segmentada, priorizando o tipo de cupom. Recomenda-se a adoção dos modelos de regressão específicos para Cashback, Produto e Desconto como ferramentas para o estabelecimento de metas e para a tomada de decisões estratégicas, aproveitando a alta previsibilidade dos cupons de Cashback e o elevado potencial de retorno dos cupons de Desconto.

Análise dos dados presentes em compras.csv

A planilha compras.csv apresenta duas variáveis fundamentalmente numéricas que serão a base de todas as análises realizadas daqui em diante, são elas a coluna valor_compra (que será tratada como valor da compra durante a análise), que descreve o valor da compra que o usuário fez em um dado estabelecimento, e valor_cupom (que, assim como na análise anterior, será tratado como valor do cupom), que descreve o valor que o cupom abateu no total da compra, os cupons novamente estão agrupados como cashback, produto e desconto que, novamente, serão fundamentais nas análises mais minuciosas a serem realizadas. Além disso, a tabela vai apresentar outras oito colunas não numéricas sendo elas numero_celular, data_captura, tipo_loja, local_captura, latitude, longitude, nome_loja e endereco_loja. Para a nossa análise, serão consideradas as variáveis data de captura e tipo da loja, pelo fato das outras ocasionarem muitos subgrupos, tornando a análise muito extensa, que possuem valor semelhante a ramificação mais curta por tipo de loja.

Optei por segmentar os dados por tipo_cupom (Cashback, Produto e Desconto), uma metodologia já utilizada na análise anterior. Em seguida, criei variáveis para armazenar a soma dos 'valores de cupom' e a soma dos 'valores de compra' para cada um desses tipos, permitindo uma análise agregada por categoria.

Após isso, ao identificar incoerências entre o nome da loja e o tipo da loja como na linha nove, onde Livraria Cultura é listada como restaurante. Para corrigir essa falha na listagem, eu optei, assim como na análise anterior, de, por meio do Python, utilizar uma função de dicionário para renomear todas as lojas listadas de acordo com o tipo legítimo da loja. Dessa forma Livraria Cultura foi corretamente listada como papelaria. Essa mesma correção foi feita para todas as lojas presentes na planilha, que foram corretamente listadas, a fim de obter resultados mais próximos da realidade e palpáveis.

Por último, as datas foram substituídas pelos seus respectivos dias da semana, com a finalidade de proporcionar uma categorização mais precisa, resumida e padronizada.

Os gráficos serão gerados por meio das bibliotecas matplotlib.pyplot e seaborn, popularmente descritas na codificação como plt e sns, respectivamente. Caso se faça necessária a geração de tabelas, essas serão definidas diretamente pelo Word.

Comparação entre todos os valores de cupons e valores de compra:

Para realizar todas as análises da planilha compras.csv, utilizarei o valor dos cupons como variável dependente (y) e o valor das compras como variável independente (x), dessa forma todas as análises preditivas realizadas utilizarão o valor das compras para prever o possível valor de cupom que será utilizado pelo usuário e se é, de fato, uma possibilidade válida e estratégica.

Análise de valor de compra x valor de cupom

$\boldsymbol{r = 0.3747}$ ...

Como r → 0,3747, utilizaremos 0,3747 para descrever e representar o coeficiente de correlação

Com esse valor, também poderemos também inferir o valor de R^2, que será calculado pelo valor de r elevado ao quadrado, dessa forma obtemos que:

$\boldsymbol{R^2 = (0,3747)^2 = 0,14041}$

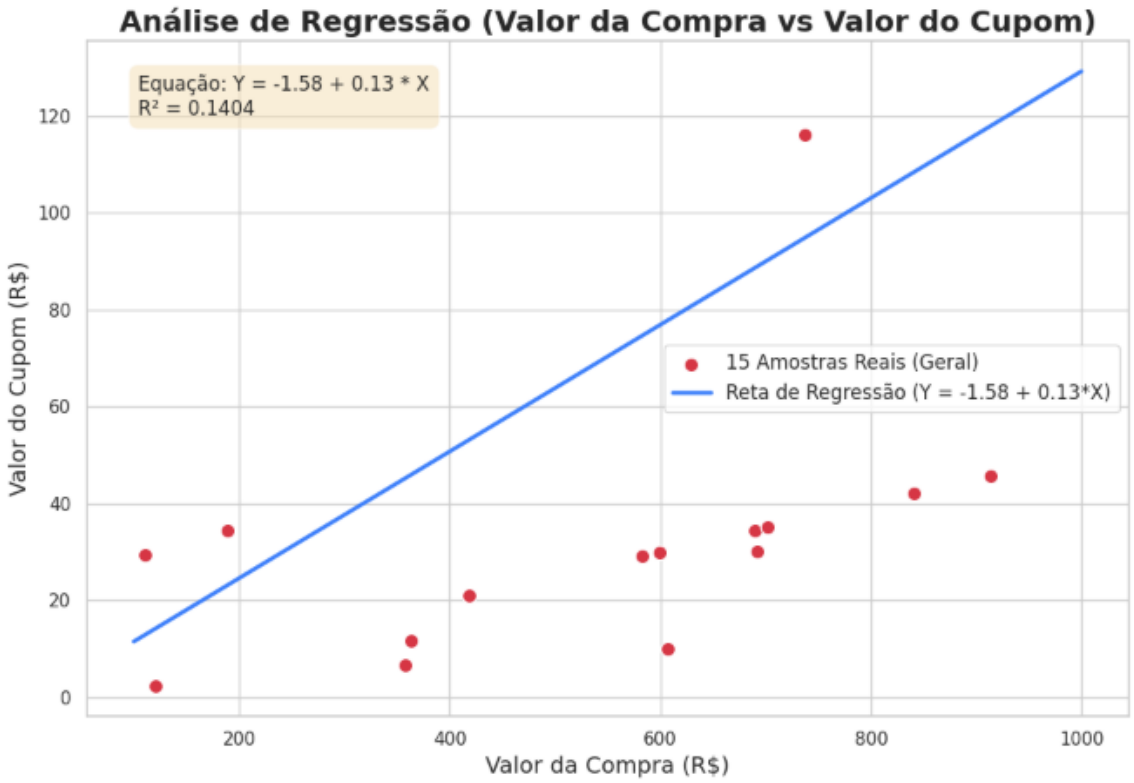
Agora para finalizar a análise numérica, vamos calcular a regressão desse conjunto de dados, de modo a obter que:

$\boldsymbol{y = -1,58 + 0,1309x}$

Os resultados obtidos anteriormente indicam a existência de uma correlação linear positiva, porém fraca, entre as duas variáveis, com um coeficiente (r) de 0,3747. Essa noção é corroborada também pelo coeficiente de determinação (R^2) de 0,14041, demonstrando que aproximadamente 14,04% da variabilidade no valor do cupom por compra feita utilizando cupons da PicMoney pode ser atrelado à variação do valor da compra, o que é um valor muito baixo e pouco determinante.

A equação de regressão obtida, demonstrada $y = -1,58 + 0,1309x$ sugere que para cada R$ do valor da compra, há um acréscimo de 0,1309R$ no valor do cupom, começando a partir de -1,58R$. No entanto, apesar de ter um certo valor preditivo, por conta da baixa correlação, muitos dos valores testados empiricamente pela equação acima vão acabar estando em dissonância do valor real, como será demonstrada pelo gráfico a seguir:

Os pontos presentes no gráfico representam em exemplar real presente na planilha. Representando um par exato entre valor da compra (eixo x) e o valor do cupom correspondente (eixo y). Para não poluir o gráfico, foram sorteadas apenas 15 amostras para serem exibidas. A reta de regressão representa qual deveria ser o posicionamento desses pontos, caso houvesse uma correlação perfeita (1 para 1).



Dessa forma, de fato, o poder preditivo do valor do cupom para valores de compra generalizados acaba sendo bem limitado, por conta da baixa correlação direta entre essas variáveis. O que é reforçado pela tendência de distância entre os pontos sortidos e a reta de regressão.

Análise de valor de compra x valor de cupom por tipo de cupom

Como a análise feita anteriormente não indica índices muito práticos e preditivos, realizaremos a mesma análise feita anteriormente, mas ramificada por tipos diferentes de cupons, com a finalidade de obter variáveis com um ajuste de maior qualidade, e, portanto, de maior potencial avaliativo. Analisarei, respectivamente, Cashback, Produto e Desconto.

Análise de valor cupom x valor de compra (Cashback)

$\boldsymbol{r = 0,9999}$ ...

Como, $r \rightarrow 1$, utilizaremos 1 para descrever e representar o coeficiente de correlação.

Com esse valor, também poderemos também inferir o valor de R^2, que será calculado pelo valor de r elevado ao quadrado, dessa forma obtemos que:

$$\boldsymbol{R^2 = 1^2 = 1}$$

Agora para finalizar a análise numérica, vamos calcular a regressão desse conjunto de dados, de modo a obter que:

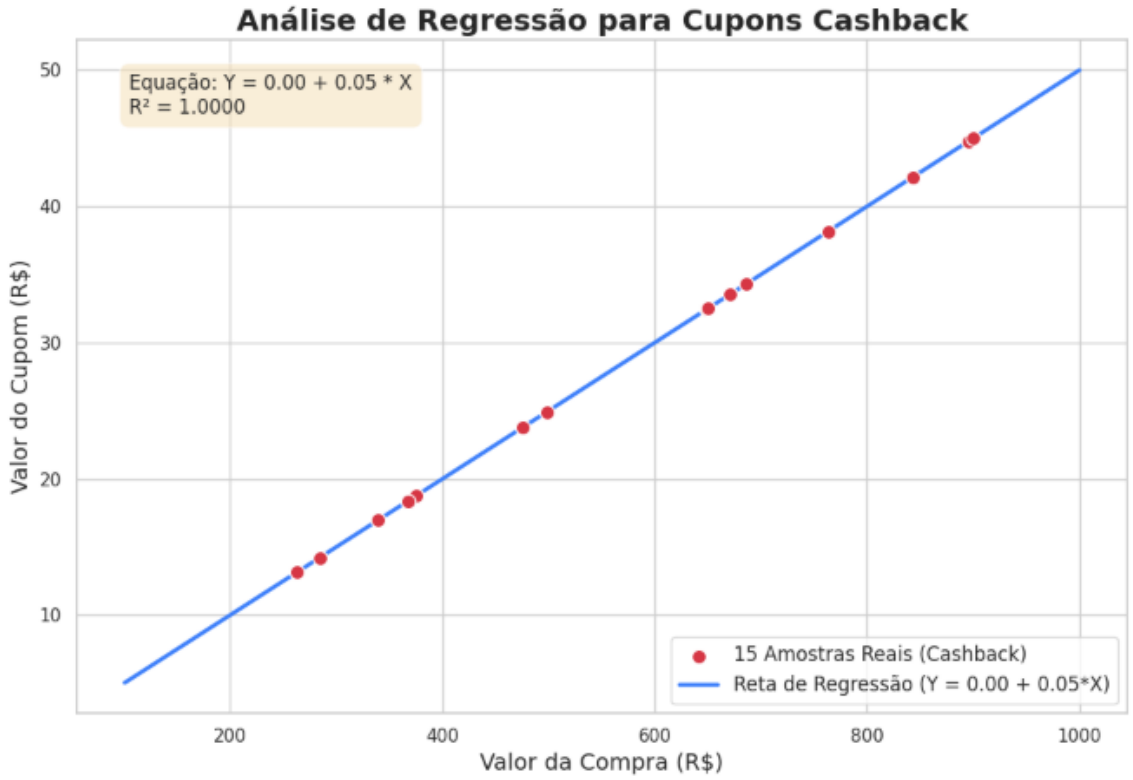
$$\boldsymbol{y = 0,05x}$$

Os resultados obtidos indicam uma correlação linear positiva, praticamente perfeita entre as duas variáveis, com um coeficiente (r) que tende consideravelmente a 1, que foi o valor utilizado para a análise. Tal concepção é também corroborada pelo coeficiente de determinação (R^2), demonstrando que 100% da variabilidade no valor do cupom por cupom de Cashback utilizado pode ser atrelada à variação do valor da compra.

A equação de regressão obtida, demonstrada $y = 0{,}05x$ sugere que para cada R$ acrescido ao valor da compra, há um acréscimo de 0,05R$ no valor do cupom. Devido à correlação de um para a um, todos os valores testados empiricamente pela equação acima vão acabar estando em conluio com o valor real, como será demonstrado pelo exemplo e pelo gráfico a seguir:

Linha oito da tabela compras.csv (id_cupom = CUP542835) tem um cupom cashback de valor R$ 534,74 para uma compra feita na Daiso Japan, no dia 26/07/2025.

$y = 534{,}74 * 0{,}05 = 26.74$, o mesmo valor de cupom presente na linha oito.



A confluência entre os pontos de amostras exibidos no gráfico e a reta de regressão confirmam a tendência elevadíssima de correlação entre o valor de cupom e o valor de compra.

Análise de valor de cupom x valor de compra (Produto)

$\boldsymbol{r = 0,7319}$ ...

Como, $\mathrm{r} \rightarrow 0{,}732$, utilizaremos 0,732 para descrever e representar o coeficiente de correlação.

Com esse valor, também poderemos também inferir o valor de R^2, que será calculado pelo valor de r elevado ao quadrado, dessa forma obtemos que:

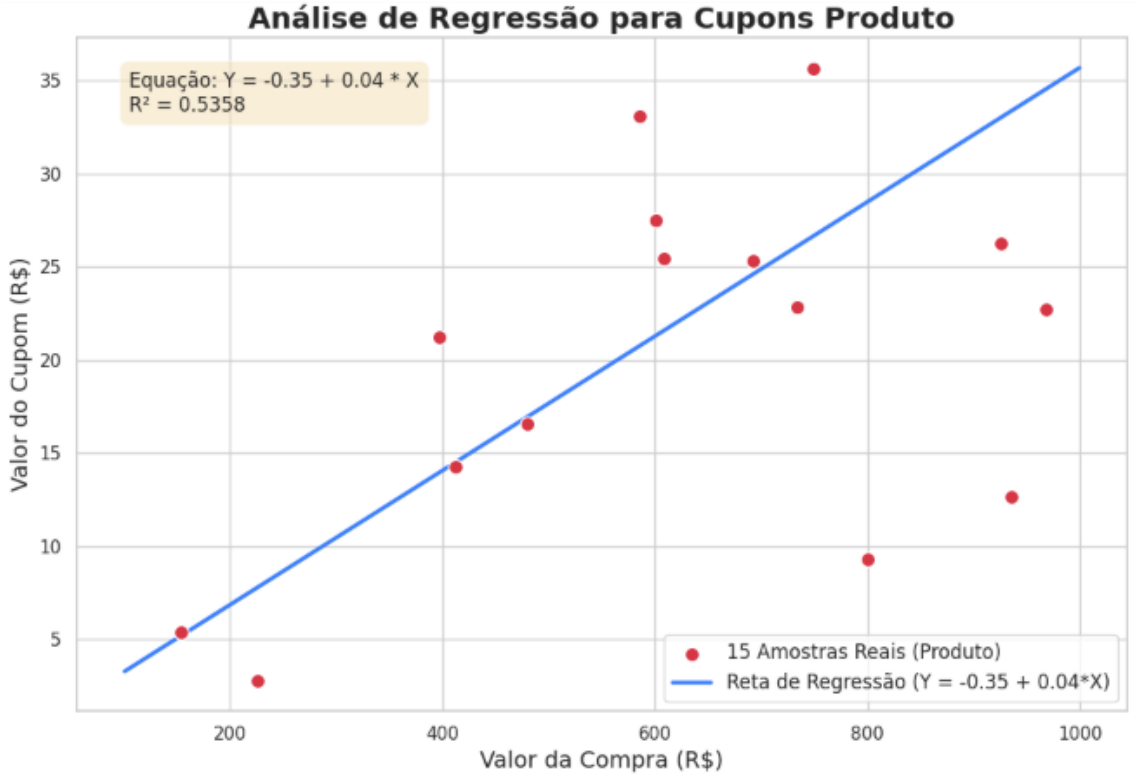
$$\boldsymbol{R^2 = 0,732^2 = 0,535824}$$

Agora para finalizar a análise numérica, vamos calcular a regressão desse conjunto de dados, de modo a obter que:

$$\boldsymbol{y = -0,35 + 0,036x}$$

Ao segmentar a análise para os cupons do tipo Produto, os resultados obtidos indicam a existência de uma correlação linear positiva forte entre as duas variáveis, com um coeficiente (r) de 0,732. Essa noção é corroborada também pelo coeficiente de determinação (R^2) de 0,535824, demonstrando que 53,5824% da variabilidade no valor do cupom pode ser atrelada à variação do valor da compra utilizando este segmento de cupons específico.

A equação de regressão obtida, demonstrada $y = -0{,}35 + 0{,}036x$ sugere que para cada R$ acrescido ao valor do cupom, há um acréscimo de 0,036R$ no repasse para a PicMoney, começando a partir de -0,35R$. Ao contrário da análise generalizada, que possuía uma correlação fraca, o valor preditivo do modelo para cupons de produto é consideravelmente mais robusto. Por conta da forte correlação, os valores previstos pela equação de regressão tenderiam a apresentar maior consonância com os valores reais, tornando o valor do cupom um indicador mais confiável e útil para prever o valor do repasse, como demonstrado pelo gráfico a seguir:



Como demonstrado no gráfico, há um poder preditivo consideravelmente maior para a análise do valor de compra em cupons de produto, em comparação com o modelo generalizado. A menor dispersão das amostras em torno da linha de regressão evidencia uma relação muito mais consistente e previsível, apesar de não ser tão precisa quanto o modelo de Cashback, como pode ser visto pelo fato de algumas amostras estarem muito próximas da reta regressiva, enquanto algumas estão um pouco mais distantes.

Portanto, trata-se de um modelo com uma boa relevância estratégica, que pode ser utilizado para auxiliar na definição de parcerias, demonstrando como o valor dos cupons de produto poderiam influenciar nos gastos dos clientes.

Análise de valor de cupom x valor de compra (Desconto)

$\boldsymbol{r = 0,7506}$ ...

Como, $\mathrm{r} \rightarrow 0{,}751$, utilizaremos 0,751 para descrever e representar o coeficiente de correlação.

Com esse valor, também poderemos também inferir o valor de R^2, que será calculado pelo valor de r elevado ao quadrado, dessa forma obtemos que:

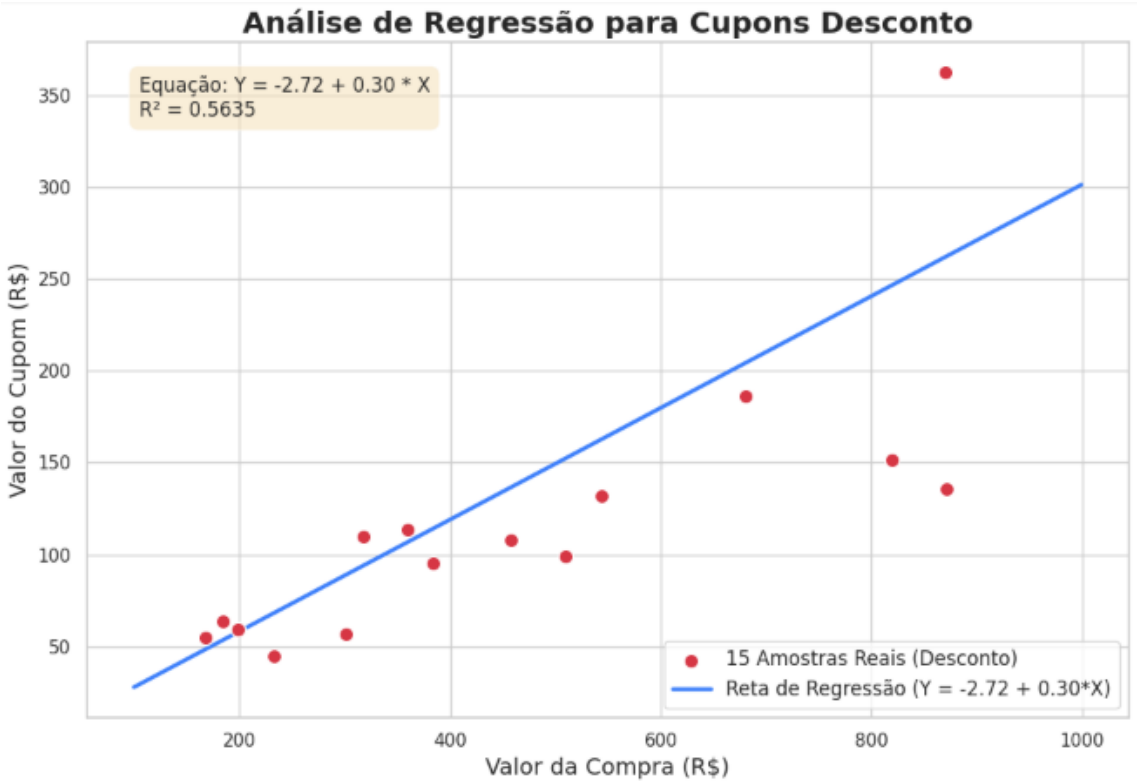
$$\boldsymbol{R^2 = 0,751^2 = 0,5635}$$

Agora para finalizar a análise numérica, vamos calcular a regressão desse conjunto de dados, de modo a obter que:

$$y = -2,72 + 0,3045x$$

Focando nos cupons do tipo Desconto, os resultados revelam uma robusta conexão linear positiva entre as variáveis, quantificada por um coeficiente de correlação (r) de 0,751. O coeficiente de determinação (R^2) de 0,564 reforça essa descoberta, explicitando que 56,4% das flutuações no valor do cupom são explicadas pela variação no valor da compra para esta categoria.

O modelo de regressão (y= −2,72 + 0,3045x) traduz essa relação em termos práticos: para cada real gasto na compra, o valor do cupom tende a aumentar em R$ 0,3045. Este resultado representa um avanço significativo em relação ao modelo generalizado, cujo baixo poder preditivo o tornava inviável. A força desta correlação valida o uso do valor da compra como um indicador confiável para estimar o valor dos cupons de desconto, como será ilustrado no gráfico correspondente:



Como demonstrado no gráfico, a análise para cupons de desconto revela um poder preditivo robusto, superior ao do modelo generalizado e muito parecido com o modelo de produto, mas com uma inclinação consideravelmente maior (Aproximadamente 30 centavos contra 4 centavos). A dispersão dos pontos, embora presente, é visivelmente menor em comparação com o modelo geral, indicando uma relação mais estável e previsível entre o valor da compra e o valor do cupom. Ainda que não alcance a precisão determinística do modelo de Cashback, a tendência é clara, com a maioria das amostras se alinhando próximas à reta de regressão.

Dessa forma, este é um modelo de alta relevância estratégica, cujo principal valor está em demonstrar a alta taxa de retorno em cupons gerados a partir do valor gasto pelo cliente. Com uma inclinação acentuada (0,30), o modelo mostra que as campanhas de desconto geram um valor de cupom significativamente maior por real gasto em comparação a outras modalidades. Esta informação é vital para o planejamento de marketing e para o controle orçamentário de futuras promoções.

Análise de valor de cupom x valor de compra por dia da semana:

Dando sequência à análise, avaliarei agora o coeficiente de correlação, o coeficiente de determinação e a regressão, mas considerando a segmentação pelos diferentes dias da semana, que foram definidos na etapa de padronização do código em Python.

Esta análise busca verificar se, de forma análoga ao tipo de cupom, o dia da semana em que a compra ocorre impacta a relação entre o valor da compra e o valor do cupom gerado, visando encontrar padrões que possibilitem uma maior capacidade de análise inferencial e preditiva. Demonstrarei os dados por meio de uma tabela, pois ao se realizar o cálculo, percebe-se um padrão muito semelhante ao encontrado na análise geral.

	Coeficiente de correlação (r)	Coeficiente de determinação (R^2)	Regressão Linear
Domingo	0,3940	0,155236	$y = -5,48 + 0,1398x$
Segunda-feira	0,3596	0,129312	$y = 14,53 + 0,0944x$
Terça-feira	0,4295	0,184470	$y = 5,63 + 0,1081x$
Quarta-feira	0,2395	0,057360	$y = -2,44 + 0,1333x$
Quinta-feira	0,3294	0,108504	$y = 2,26 + 0,1272x$
Sexta-feira	0,3656	0,133663	$y = -0,25 + 0,1276x$
Sábado	0,3674	0,134997	$y = -9,11 + 0,1547x$

Os dados obtidos na tabela acima apresentam um coeficiente de correlação de Pearson bastante variado, oscilando de um pico de 0,4295 na terça feira, para um baixo de 0,2395 na quarta-feira. No entanto, mesmo considerando o coeficiente mais elevado, apresenta um coeficiente de determinação baixíssimo na casa de aproximadamente 18,5%, enquanto o coeficiente mais baixo apresenta um coeficiente de determinação de apenas 5,7%, contemplando um valor preditivo praticamente ínfimo.

A regressão também indica uma contradição que vale ser ressaltada, por exemplo, no sábado, o dia que apresenta a maior inclinação, de pouco mais de 0,15 centavos, mesmo não sendo o dia com a maior correlação. Enquanto o dia com a maior correlação, apresenta uma inclinação de aproximadamente 0,11 centavos, valor inferior ao do dia com a menor correlação.

Portanto, a correlação segmentada por dia da semana, além de ter um baixíssimo valor preditivo, apresenta uma inconsistência que torna o seu uso ainda mais insensato e ineficaz. É um valor que se aproxima bastante da taxa geral, possuindo em alguns dias valores de correlação que superam o do modelo geral, e outros que ficam muito abaixo do modelo geral.

Análise de valor de cupom x valor de compra por tipo de loja:

Agora a última variável a ser analisada será o tipo de loja, avaliarei o coeficiente de correlação, o coeficiente de determinação e a regressão, como fator crucial da análise.

Esta análise busca verificar se, de forma análoga ao tipo de cupom, o tipo em que a compra ocorre impacta a relação entre o valor da compra e o valor do cupom gerado, visando encontrar padrões que possibilitem uma maior capacidade de análise inferencial e preditiva. Essa análise também busca entender se o tipo de loja apresenta um impacto maior do que o dia da semana. Novamente, os dados serão separados por meio de uma tabela.

	Coeficiente de correlação (r)	Coeficiente de determinação (R^2)	Regressão Linear
Eletrodoméstico	0,3816	0,145618	$y = 0,04 + 0,126x$
Esportivo	0,3914	0,153193	$y = -4,29 + 0,1346x$
Farmácia	0,3621	0,131116	$y = 2,86 + 0,1267x$
Mercado Express	0,3810	0,145161	$y = -6,52 + 0,1367x$
Móveis	0,3757	0,141150	$y = -3,44 + 0,1393x$
Outros	0,3772	0,142279	$y = -3,37 + 0,1289$
Restaurante	0,3820	0,145924	$y = -1,6 + 0,1401x$
Vestuário	0,3492	0,121940	$y = 2,99 + 0,1160x$

Os dados obtidos na tabela acima apresentam um coeficiente de correlação de Pearson bastante consistente, oscilando de um pico de $0,3914$ para lojas esportivas, para um baixo de $0,3492$ para lojas de vestuário. São valores que se aproximam bastante do valor geral de $0,3747$.

A regressão, apesar de diferenças no intercepto, apresenta valores de inclinação muito aproximados, indicando que, de fato o tipo de loja é pouco influente para a elaboração de qualquer estratégia financeira.

Dessa forma, torna-se muito viável afirmar que o tipo de loja tem uma influência ínfima na determinação do preço de cupom pelo preço da compra.

Conclusão sobre compras.csv:

A análise inferencial realizada sobre a base de dados cupons.csv teve como objetivo compreender a relação entre o valor da compra realizada pelos usuários e o respectivo valor de cupom que o usuário usaria, a fim de criar

modelos preditivos para a estratégia da empresa. Principalmente na elaboração de parcerias e quantificação preditiva da distribuição de cupons.

A análise inicial demonstrou que um modelo generalizado, considerando o volume total de cupons, é bastante ineficaz e impreciso. A correlação geral entre o valor da compra e o valor do cupom mostrou-se fraca ($r \approx 0{,}37$) e com um poder preditivo muito baixo ($R^2 \approx 0{,}14$), sendo inadequado para decisões estratégicas. Da mesma forma, a segmentação por dia da semana também se mostrou infrutífera e inconsistente, pois a relação entre as variáveis permaneceu fraca e imprevisível em todos os dias da semana. Além disso, a segmentação por tipo de loja, também se mostrou bastante fútil, pois retornou índices muito semelhantes aos calculados inicialmente na análise generalizada, destacando que o tipo de loja tem uma influência ínfima na definição do valor dos cupons.

O principal modelo preditivo definido pela análise foi, assim como na planilha analisada anteriormente, a segmentação pela variável de tipo do cupom, que se revelou, novamente, como o único fator que determinadamente influência no comportamento para a determinação do valor do cupom se comparado com o valor da compra. Podendo concluir-se que:

1. Cashback: Apresenta uma correlação perfeita ($r = 1{,}0$), operando como uma regra de negócio fixa e 100% previsível, onde o valor do cupom corresponde a exatamente 5% do valor da compra.

2. Desconto: Possui uma correlação forte ($r \approx 0{,}75$) e um bom poder preditivo ($R^2 \approx 0{,}56$). Sua principal característica é a alta taxa de retorno, onde cada $R\$\ 1{,}00$ gasto na compra gera aproximadamente $R\$\ 0{,}30$ em valor de cupom.

3. Produto: Também com correlação forte ($r \approx 0{,}73$) e bom poder preditivo ($R^2 \approx 0{,}54$), este modelo opera com uma taxa de retorno drasticamente inferior, gerando apenas cerca de $R\$\ 0{,}036$ em cupom para cada $R\$\ 1{,}00$ gasto.

Do ponto de vista estratégico, esta distinção é fundamental. Para a elaboração de parcerias, fica claro que um parceiro que utiliza cupons de "Desconto" terá um impacto financeiro (custo em cupons) quase 8.5 vezes maior do que um que utiliza cupons de "Produto". Na quantificação preditiva, os modelos de Desconto e Produto oferecem uma base confiável (com mais de 50% de acerto) para estimar os valores de cupons a serem distribuídos, enquanto o modelo de Cashback permite um cálculo exato.

Conclui-se, portanto, que, dadas as variáveis presentes na planilha, a análise segmentada por tipo_cupom é a única abordagem válida para a tomada de decisões, descartando o modelo geral e a segmentação por dia da semana. Vale-se destacar que com a consideração e disponibilidade de outras variáveis, pode-se conquistar fatores que tenham uma influência considerável nessa

análise. Portanto, com o que fora analisado, a empresa dispõe de três ferramentas distintas que possuem um elevado valor preditivo.

Análise dos dados presentes em pedestres.csv

A planilha pedestres.csv apresenta somente duas variáveis essencialmente numérica que são a coluna ultimo_valor_capturado (que daqui em diante será tratada como último valor capturado) e idade. Além disso, outras colunas presentes na tabela são horário, data, local, tipo_celular (se refere à marca do celular), modelo_celular, possui_app e sexo. A planilha, da maneira que está no momento, possui pouquíssimas possibilidades de inferência de dados, de modo que tende a ser necessária a quantificação de uns dados e substituição por um coeficiente e representante numérico, a fim de obter dados de correlação, regressão e determinação palatáveis.

Essa análise será mais curta, pois não necessitará de muitas segmentações e os resultados seguirão um padrão pouco conclusivo. Nessa análise verificaremos se há alguma relação entre o último valor resgatado por um usuário e a sua idade, a fim de entender se usuários de diferentes grupos etários tendem a utilizar cupons mais baratos ou mais caros.

Além disso, não houve a necessidade de nenhum Data Cleansing, pois não identifiquei inconsistências nos dados e a estruturação já era satisfatória para as análises que seriam feitas. Nesse excerto, último valor capturado será a variável dependente (y), enquanto a idade será a variável independente (x). Será feita a análise por idade e último valor de cupom e, também, serão consideras as variáveis não numéricas de tipo de cupom, tipo de celular (sistema operacional), sexo, tipo da loja

Comparação entre todas as idades e últimos valores de cupom:

$$\boldsymbol{r = 0,003}$$

Com esse valor, também poderemos também inferir o valor de R^2, que será calculado pelo valor de r elevado ao quadrado, dessa forma obtemos que:

$$\boldsymbol{R^2 = (0,003)^2 = 0,000009}$$

Agora para finalizar a análise numérica, vamos calcular a regressão desse conjunto de dados, de modo a obter que:

$$\boldsymbol{y = 253,62 + 0,0272x}$$

Os resultados obtidos anteriormente indicam, praticamente, a inexistência de uma correlação linear entre as duas variáveis, com um coeficiente (r) de 0,003, que deriva um coeficiente determinação (R^2) que tende consideravelmente a 0, demonstrando que apenas $0{,}0009\%$ da variabilidade no valor de cupom retirado pode ser atrelada à variação da idade do usuário

Conclusão sobre pedestres.csv

Os fatores que foram na análise feita anteriormente tornam uma demonstração de regressão ou uma demonstração gráfica consideravelmente inviáveis, pois teriam um valor analítico e preditivo praticamente nulo, de modo que não poderiam ajudar nas estratégias financeiras de maneira alguma. Além disso, possíveis segmentações, seguindo outras variáveis tais quais o dia da semana, o tipo do celular, o tipo da loja ou sexo do usuário pouco valor acrescentariam a essa análise.

Análise dos dados presentes em players.csv

A planilha players.csv, que trata de informações demográficas de alguns usuários, como idade, data de nascimento, bairro/cidade residencial, bairro/cidade em que estuda, bairro/cidade em que trabalha, número de celular e categoria que costuma frequentar, dos dados ordenados anteriormente, somente idade é um dado numéricas, enquanto os demais tratam de dados não numéricos e de pouco valor analítico, especialmente se o objetivo é analisar fatores como a correlação e a regressão desses dados. Por esse motivo o conjunto de dados presentes em players.csv não seguirá o padrão de análise utilizado nas demais planilhas, pois não dispõe de uma base analítica sólida.

Conclusão geral:

A análise inferencial conduzida sobre as quatro planilhas disponibilizadas pela PicMoney permitiu avaliar, de forma segmentada, a relevância estatística de diferentes variáveis na construção de modelos preditivos.

No caso de cupons.csv, observou-se que a relação geral entre valor do cupom e repasse à PicMoney é fraca ($r \approx 0{,}37$; $R^2 \approx 0{,}13$), se tratando de um modelo pouco viável. Variáveis temporais (período do dia, dia da semana) e categóricas (tipo de estabelecimento) também não acrescentaram um valor preditivo relevante. O fator decisivo foi o tipo de cupom, destacando-se Cashback (correlação praticamente perfeita, $r \approx 1{,}0$; $R^2 = 1{,}0$), seguido de Produto ($r \approx 0{,}72$; $R^2 \approx 0{,}52$) e Desconto ($r \approx 0{,}74$; $R^2 \approx 0{,}55$). Conclui-se, portanto, que apenas a segmentação por tipo de cupom possui uma relevância estratégica e avaliativa nesta base.

A análise de compras.csv apresentou resultados análogos. A relação entre valor de compra e valor de cupom mostrou-se igualmente fraca em termos gerais ($r \approx 0{,}37$; $R^2 \approx 0{,}14$), inclusive retornando valores muito parecidos com a primeira planilha, e segmentações por dia da semana ou tipo de loja revelaram baixa capacidade analítica. Novamente, o tipo de cupom se destacou como a variável mais relevante, Cashback apresentou uma correlação altíssima ($r \approx 1{,}0$), Produto demonstrou correlação forte ($r \approx 0{,}73$; $R^2 \approx 0{,}54$), e Desconto também se destacou ($r \approx 0{,}75$; $R^2 \approx 0{,}56$), com maior inclinação regressiva, indicando um impacto financeiro maior. Esta diferenciação torna o tipo de cupom o único parâmetro válido para projeções confiáveis nesta planilha.

Já em pedestres.csv, a relação entre idade e último valor capturado apresentou correlação praticamente nula ($r \approx 0{,}003$; $R^2 \approx 0{,}0009$), descartando qualquer aplicabilidade prática. Segmentações adicionais, considerando variáveis como tipo de celular, sexo ou tipo de loja, não se mostraram relevantes e não alteraram o caráter não preditivo dos dados.

Por fim, players.csv revelou-se a base menos adequada para análises inferenciais. A predominância de variáveis não numéricas inviabilizou cálculos de correlação ou regressão com significado estatístico, restringindo sua utilidade ao caráter descritivo e cadastral.

De forma concisa, conclui-se que os únicos modelos estatisticamente válidos e de real valor estratégico derivam da segmentação por tipo de cupom, especialmente nas bases cupons.csv e compras.csv. Cashback opera como modelo determinístico e mais previsível, enquanto Produto e Desconto apresentam forte poder preditivo, com destaque para o maior impacto financeiro dos cupons de Desconto. Já pedestres.csv e players.csv não forneceram padrões relevantes para previsão ou suporte à tomada de decisão. Assim,

recomenda-se que a PicMoney direcione sua modelagem preditiva e estratégias de negócio exclusivamente aos resultados segmentados por tipo de cupom e, se possível, disponibilizar ou analisar mais fatores que podem acabar auxiliando na compreensão das diferenças análises de correlação e regressão.