Projeto Interdisciplinar: Ciências de Dados
Entrega 1

Guilherme Barioni RA:24026140
Iury Xavier da Silva Mangueira RA:24026311
Lilian Mercedes Paye Conde RA:24026462
Marcus Miranda Duque RA:24026080
Murilo de Souza Vieira RA:24025726

Base de transações – Cupons Capturados

Descrever os Dados:

A tabela contém os dados de cupons capturados e detalhes relacionados ao Player e o preço dos cupons, os dados são:

- **Celular:** O numero do celular do Player que realizou a captura.
- **Data:** O dia, mês e ano da captura do cupom.
- **Hora:** O horário da captura, no formato hora, minuto e segundo.
- **Nome_Estabelecimento:** Nome do estabelecimento onde o cupom foi capturado.
- **Bairro_ Estabelecimento:** Bairro onde o estabelecimento é localizado.
- **Categoria_Estabelecimento:** Categorias de negócios que correspondem as atividades do estabelecimento.
- **ID_Campanha:** Identificação da campanha de distribuição dos cupons.
- **ID_Cupom:** Identificação individual do cupom capturado.
- **Tipo_Cupom:** Tipo de cupom capturados cashback, desconto ou produto.
- **Produto:** Produto comprado com o cupom resgatado.
- **Valor_Cupom:** Valor do cupom resgatado.
- **Repasse_Picmoney:** Valor do cupom repassado a PicMoney.

Verificar a qualidade dos dados:

A coluna “categoria_estabelecimento”, não condiz com a coluna “nome_estabelecimento”, com estabelecimentos como o Outback na categoria Igreja e Lojas de artigos religiosos ou Subway na categoria Fisioterapia e terapias complementares.

Ao verificar a qualidade da coluna “produto”, descobrimos que 66% da coluna está vazia e que os dados que estão presentes não fazem sentido:

Odio	212
Cum	211
Repudiandae	209
Ea	208
Fugiat	208

Pesquisando esses dados no Google descobrimos que são palavras em Latim, que possivelmente fazem parte de um lorem ipsum, feito para preencher espaços vazios com texto.

Limpeza dos dados:

Primeiramente, realocamos os estabelecimentos em suas categorias corretas, reduzindo a quantidade para 12 categorias, removendo categorias como Igrejas e Lojas de artigos

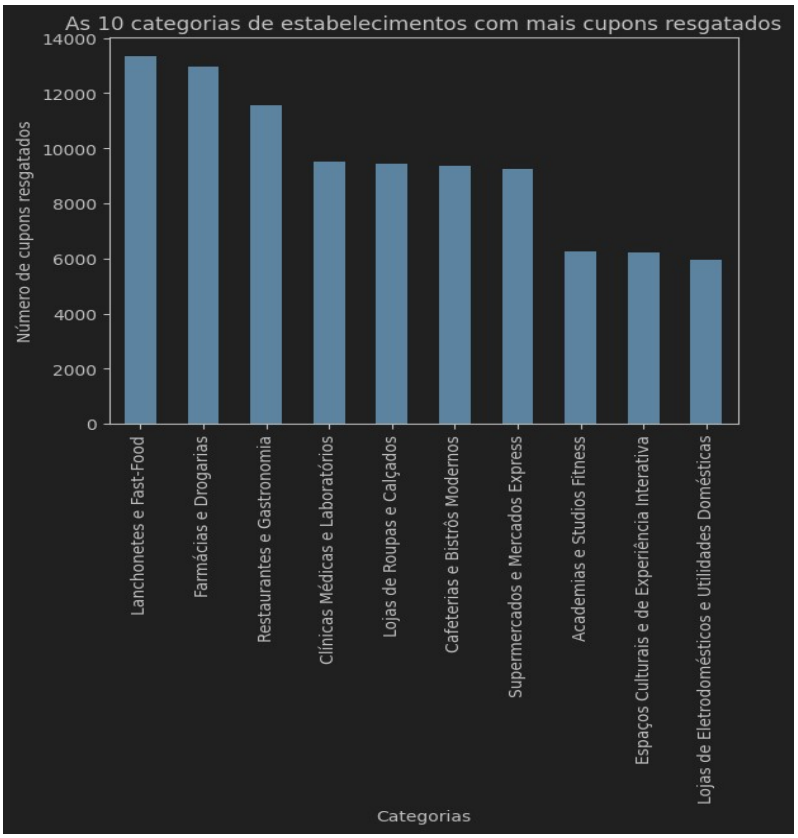
religiosos, pois não existe nenhum estabelecimento na base de dados que se encaixa nessa categoria. Com isso obtemos as seguintes categorias:

- Lanchonetes e Fast-Food
- Farmácias e Drogarias
- Restaurantes e Gastronomia
- Clínicas Médicas e Laboratórios
- Lojas de Roupas e Calçados
- Cafeterias e Bistrôs Modernos
- Supermercados e Mercados Express
- Academias e Studios Fitness
- Espaços Culturais e de Experiência Interativa
- Lojas de Eletrodomésticos e Utilidades Domésticas
- Clubes e Centros de Convivência
- Lojas de Moda Urbana e Alternativa

Para a coluna “produto”, decidimos que a melhor escolha seria removê-la, pois não contém nenhum dado que possa ser útil na nossa análise.

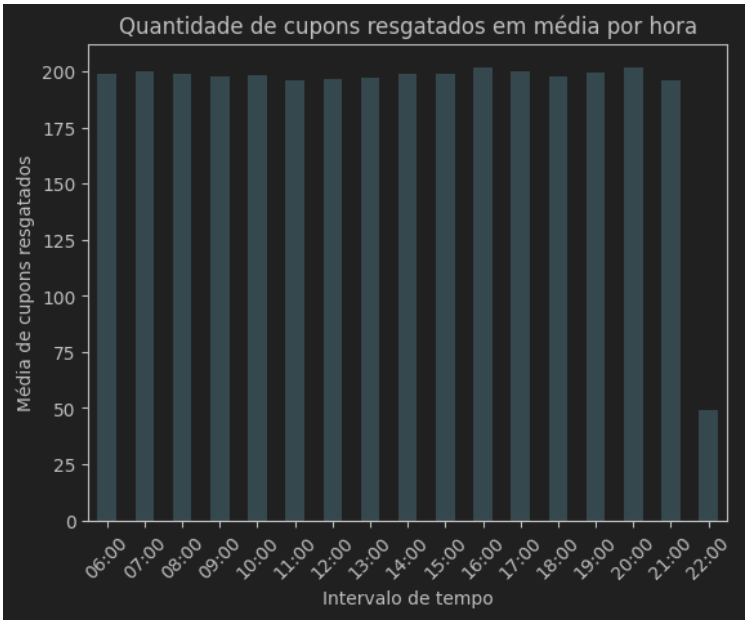
Explorar os Dados:

Com a limpeza dos dados conseguimos descobrir que as categorias mais frequentadas são:



Com Lanchonetes e Fast-Food, Farmácias e Drogarias e Restaurantes e Gastronomia sendo as mais visitadas e que mais contribuem em repasse para a PicMoney, é possível perceber que as categorias mais visitadas são de consumo cotidiano, evidenciando a utilização prática dos cupons.

Analisando a quantidade de resgates por hora é possível perceber que a maioria dos cupons são regatados durante às 06:00 até às 21:00, sendo a quantidade de capturas distribuídas igualmente durante esse horário, à também, uma forte queda de atividade às 22:00 e nenhuma atividade a partir dás 23:00 até às 05:00, com as atividades sendo resumidas às 06:00.



A PicMoney recebe em média 12,8% do valor total dos cupons como repasse.

Base Cadastral de Players

Descrever os Dados:

A tabela contém os dados cadastrados de cada player, os dados são:

- **Celular:** O número de telefone do celular cadastrado pelo player.
- **Data de nascimento:** A data de nascimento registrada por cada player, no formato dia, mês e ano.
- **Idade:** A idade de cada player.
- **Sexo:** O sexo informado pelo player.
- **Cidade_Residencial:** Cidade onde o player reside.
- **Bairro_Residencial:** Bairro onde o player reside.
- **Cidade_Trabalho:** Cidade onde o player trabalha, caso o player trabalhe.
- **Bairro_Trabalho:** Bairro onde o player trabalha, caso o player trabalhe.
- **Cidade_Escola:** Cidade onde o player estuda, caso o player estude.
- **Bairro_Escola:** Bairro onde o player estuda, caso o player estude.
- **Categoria_Estabelecimento:** Categoria de estabelecimento que o player mais frequenta.

Verificar a qualidade e Limpeza dos Dados

Ao juntar a tabela da base cadastral com a tabela da base simulada na Av. Paulista é possível perceber que a maioria dos usuários que responderam que possuem conta no app da PicMoney, não estão presentes na base cadastral, para resolver esse problema, decidimos inserir os dados desses usuários na base cadastral utilizando o número de celular, idade, sexo e categoria frequentada que foram puxados da base simulada.,

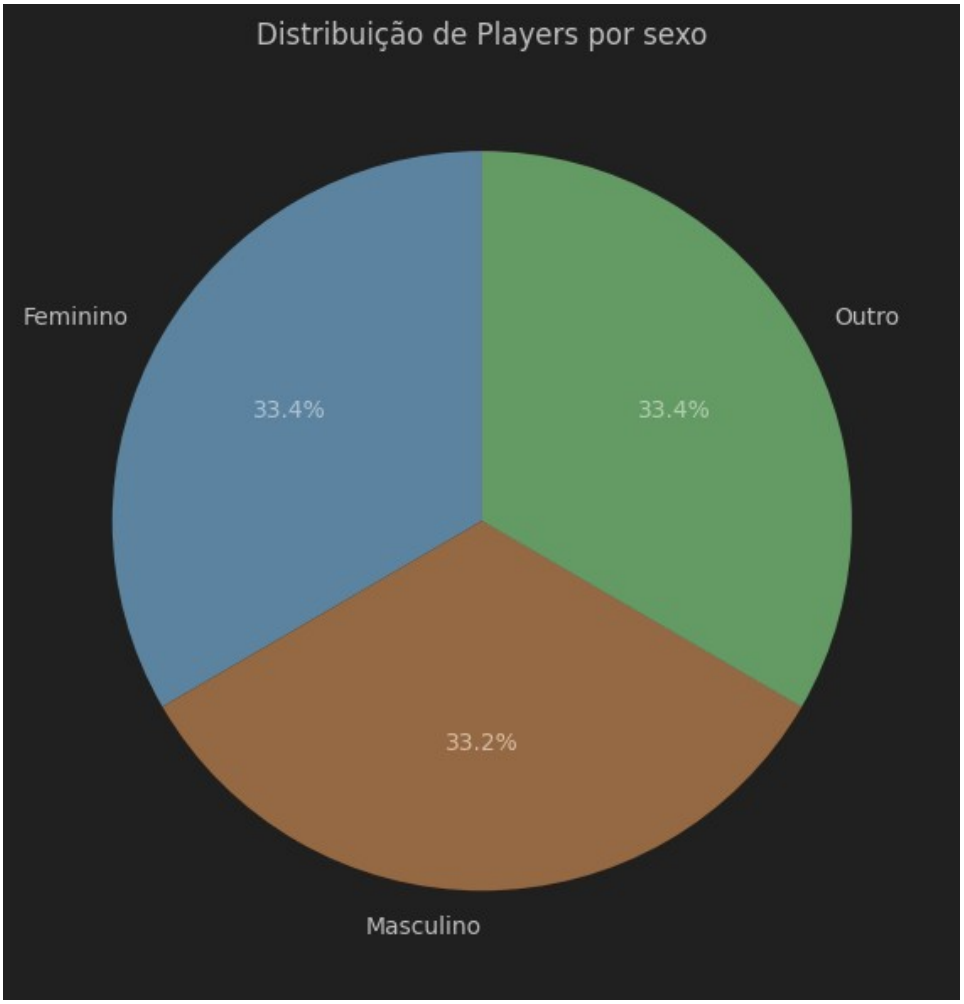
também mudamos de não para sim os usuários que estavam presentes na tabela de base cadastral e tinham respondido que não tinham o aplicativo da PicMoney.
Fizemos essa verificação com a tabela de Massa de Teste onde a maioria dos telefones que estão presentes na tabela não estão na base cadastral, adicionamos as informações com base no celular e tipo de loja, sendo as únicas informações disponíveis na massa de teste que são compatíveis com a base cadastral.

Todos os celulares presentes na tabela de cupons capturados estão na base cadastral, finalizamos excluindo todas as possíveis duplicatas, inicialmente a base cadastral tinha 10.000 usuários cadastrados e agora tem 69.949 usuários.

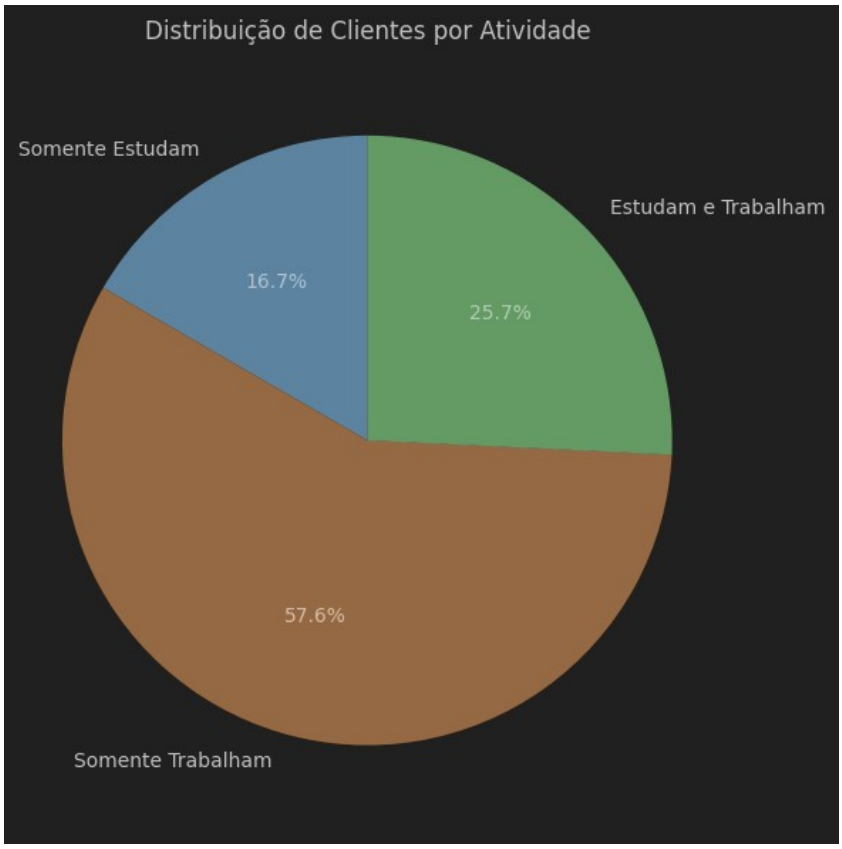
Explorando os dados

Com a base cadastral corrigida, é possível determinar que a média da idade dos usuários é de 39 anos.

A distribuição por sexo na base cadastral é praticamente igual nas três categorias, sendo em média 33,33%:



A tabela contém o local onde os usuários estudam ou trabalham, com isso é possível determinar quanto da base de usuários trabalham, estudam ou realizam os dois:



Sendo que a maioria dos usuários(83,3%) trabalham e somente 16,7% dos usuários só estudam

Base simulada de pedestres da Av. Paulista

Descrever os dados

A tabela contém os dados simulados de pedestres que resgataram cupons na avenida paulista, os dados são:

- **Celular:** O número de telefone cadastrado.
- **Data:** A data do resgate dos cupons, no formato dia, mês e ano. A tabela só inclui cupons resgatados no dia 22/07/2025.
- **Horário:** O horário em que o cupom foi capturado.
- **Local:** Descrição do local onde o cupom foi capturado.
- **Latitude:** A latitude exata de onde o cupom foi capturado.
- **Longitude:** A longitude exata de onde o cupom foi capturado.
- **Tipo_Celular:** Sistema operacional do celular do player, dois tipos estão presentes na tabela: Android ou iPhone.
- **Modelo_Celular:** Marca e modelo do celular do player.
- **possui_app_picmoney:** Marca se o player possui o app da PicMoney com um sim ou não.
- **data_ultima_compra:** Data da ultima compra realizada com a utilização da PicMoney.
- **ultimo_tipo_cupom:** Tipo do cupom utilizado na ultima compra realizada com a PicMoney, entre cashback, desconto ou produto.
- **ultimo_valor_capturado:** Valor do ultimo cupom capturado utilizando o app da PicMoney.

- **ultimo_tipo_loja:** Tipo de loja onde a ultima compra foi realizada utilizando um cupom da PicMoney.
- **Idade:** Idade do player.
- **Sexo:** Sexo do player.

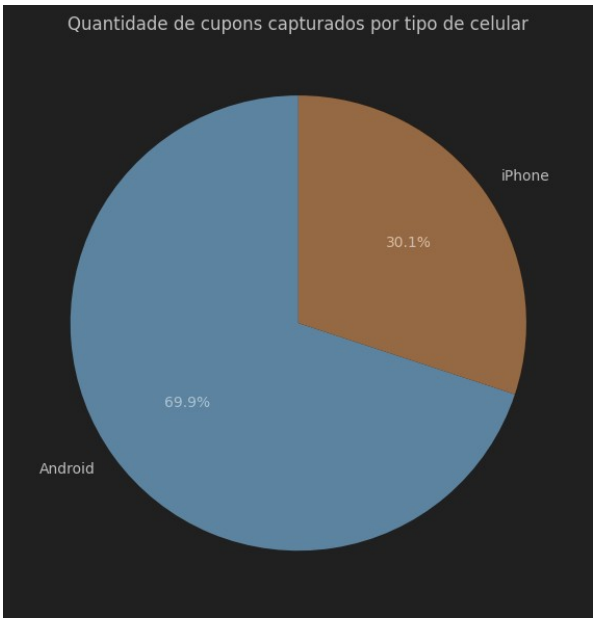
Verificar a qualidade e Limpeza dos Dados

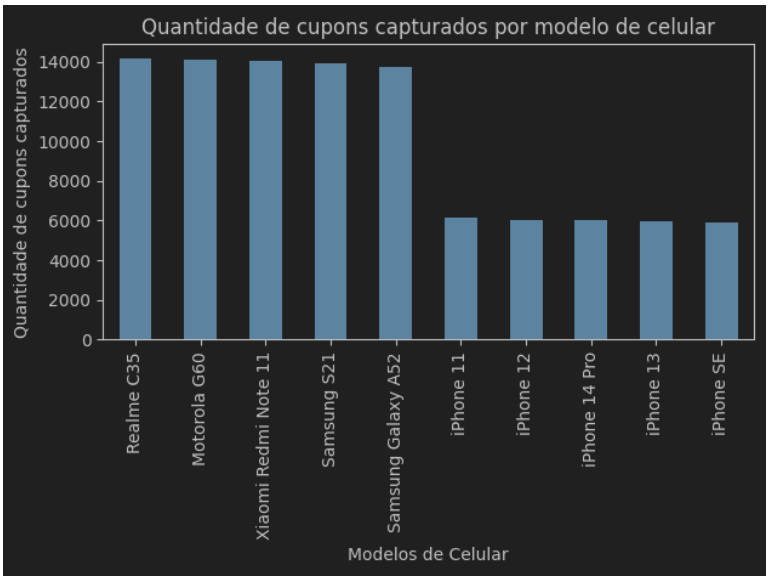
As coordenadas presentes nas colunas latitude e longitude não estão em um formato correto, para isso, removemos todos os pontos e virgulas e convertemos as coordenadas utilizando uma função que às divide por 10 múltiplas vezes até o valor está dentro da faixa real de uma coordenada.

Decidimos padronizar os dados na coluna “ultimo_tipo_loja” com as categorias de lojas presentes na coluna “tipo_loja” da tabela da Base de Transações e Cupons Capturados.

Explorando os dados

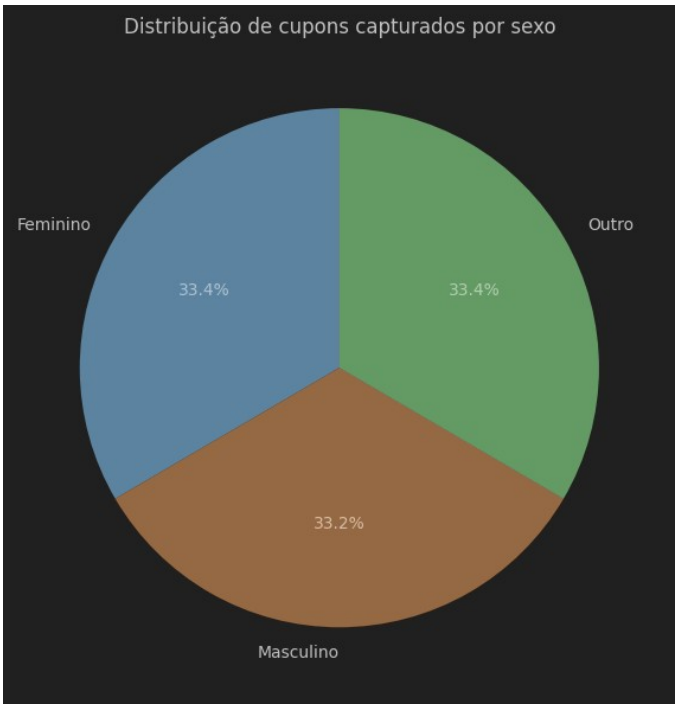
O celular mais utilizado pelos usuários são o Android com 69,9% dos usuários seguido pelo iPhone com 30,1% de usuários.





A distribuição por modelos de celular condiz com o tipo de celular mais utilizado, sendo as mais utilizadas marcas com sistema operacional Android seguidas por modelos da Apple.

Ao analisar a distribuição de cupons por sexo é possível perceber que o sexo masculino, feminino e outros individualmente compõem 33,33%, portanto não existe forte diferenciação entre o resgate de cupons por sexo.



A média de idade da base simulada é de 42 anos e o valor médio da última compra realizada é R$254,80.

Massa de teste com valores e lojas

Descrever os dados:

A tabela contém os dados de uma massa de teste, detalhando locais, valores de compra e os valores dos cupons, os dados são:

- **numero_celular:** O número do celular do player.
- **data_captura:** Data da captura do cupom.
- **tipo_cupom:** Tipo do cupom resgatado sendo ele cashback, produto ou desconto.
- **tipo_loja:** Tipo da loja onde o cupom foi resgatado.
- **local_captura:** Local onde o cupom foi capturado.
- **Latitude:** Latitude exata de onde o cupom foi capturado.
- **Longitude:** Longitude exata onde o cupom foi capturado.
- **nome_loja:** Nome da loja onde o cupom foi capturado.
- **endereco_loja:** Endereço da loja onde o cupom foi capturado, com a rua e Cep.
- **valor_compra:** Valor da compra feita pelo player.
- **valor_cupom:** Valor do cupom resgato que foi descontado na compra pela utilização do app da PicMoney.

Verificar a qualidade e Limpeza dos Dados

Novamente, as lojas não correspondem com as suas categorias, portanto realocamos as lojas para categorias mais apropriadas, com 7 categorias sendo elas:

- vestuário
- eletrodoméstico

- mercado express
- móveis
- outros
- restaurante
- farmácia

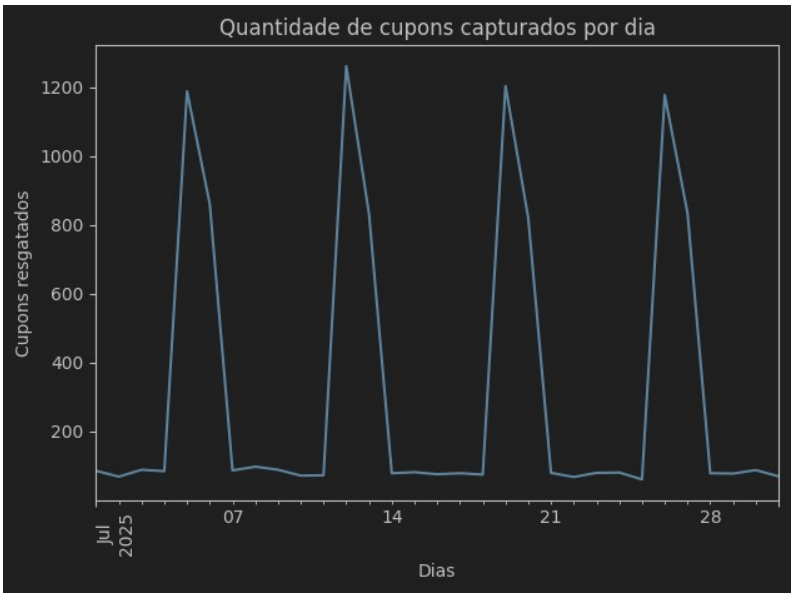
Também verificamos que a latitude e longitude presentes na tabela não estão em um formato valido, para solucionar esse problema criamos uma função que divide o valor recebido por 10, até que ele esteja dentro dentro da faixa de um valor verdadeiro para uma coordenada. Armazenamos o novo valor em duas colunas “latitude_limpa” e “longitude_limpa”.

É possível perceber que a coluna “endereco_loja” não está correta, visto que as coordenadas levam para endereços completamente diferentes, apesar de que a maioria dessas coordenadas ainda se encontram em São Paulo.

Explorar os dados

Ao analisar a quantidade de cupons capturados diariamente, é possível perceber que a maioria foi capturada durante os sábados e domingos, com uma quantidade minuscula de cupons sendo capturados entre segunda e sexta.

2025-07-12	1261
2025-07-19	1203
2025-07-05	1188
2025-07-26	1177
2025-07-06	860
2025-07-27	835
2025-07-13	832
2025-07-20	820
2025-07-08	98



Conclusão

A análise interdisciplinar dos dados permitiu identificar padrões relevantes no comportamento dos usuários da PicMoney e nas características das transações realizadas. A limpeza e padronização das bases foram etapas fundamentais para garantir a confiabilidade dos resultados, eliminando inconsistências como categorias incorretas, coordenadas inválidas e informações irrelevantes.

Com os dados tratados, foi possível observar que as categorias mais frequentadas estão relacionadas a consumo cotidiano, como lanchonetes, farmácias e restaurantes, refletindo o perfil de uso prático dos cupons. Além disso, a análise das idades e da distribuição por sexo mostrou uma base equilibrada de usuários, com média etária próxima dos 40 anos. Outro ponto de destaque foi a predominância do sistema Android entre os celulares

utilizados, além da constatação de que a maioria dos players trabalha, o que pode influenciar diretamente nos horários e locais de captura dos cupons.

Esses resultados reforçam a importância do processo de preparação de dados em projetos de ciência de dados, pois somente com uma base consistente foi possível gerar análises relevantes. A partir dessa primeira entrega, abrem-se caminhos para análises mais profundas, como estudos preditivos sobre o comportamento de consumo, segmentação de usuários e estratégias direcionadas para aumentar a eficiência da PicMoney no mercado.