Instituto de Educação Superior de Brasília - IESB
Ciência de Dados e Inteligência Artificial

Análise Exploratória de Dados do Rover Curiosity

por

Victor Augusto Souza Resende - 1922120027

Brasília - DF, 24 de Junho de 2021

Resumo

O rover Curiosity foi consequência de uma das missões mais bem-sucedidas efetuadas pelos humanos em outro planeta. Equipado com o instrumento REMS (Estação de Monitoramento Ambiental do Rover), fornece relatórios diários sobre as condições meteorológicas ao redor do rover. Com a utilização da metodologia CRISP-DM e utilização de linguagem de programação Python e SQL, foi possível analisar os dados coletados pelo instrumento REMS e fazer a previsão da temperatura máxima do solo na cratera Gale, onde o rover Curiosity encontra-se.

Palavras-chave: Curiosity; Marte; REMS; NASA; Análise de Dados, Python, SQL, CRISP-DM, Regressão-Linear

Abstract

The Curiosity rover was the result of one of the most successful missions carried out by humans on another planet. Such rover is equipped with the REMS instrument, which is the Rover's Environmental Monitoring Station, provides reports on the weather conditions around the rover. Using the CRISP-DM methodology and using the Python and SQL programming language, it was possible to analyze the data collected by the REMS instrument and make a prediction of the maximum soil temperature in the Gale crater, where the Curiosity rover is located.

Key-words: Curiosity; Mars; REMS; NASA; Data analysis; Python, SQL, CRISP-DM; Linear-Regression

Contents

List of Figures

1 Missões Nasa

A curiosidade é uma característica natural e inata do ser humano, e o que os difere dos outros animais são atributos como a racionalidade, a curiosidade e a vontade de entender o inexplorado. Estas são características que sempre estiveram presentes na história de tais. Consequentemente, com o passar dos anos, décadas e milênios, a humanidade foi capaz de começar a entender a natureza, chegando ao ponto de ser possível a criação de formulas e teoremas para a compreensão dos fenômenos da natureza. Entretanto, ainda há algo que desde o principio intriga-os, o espaço. Não há algo mais curioso que o universo.

Com o avanço armamentista e tecnológico, a National Aeronautics and Space Administration (Administração Nacional da Aeronáutica e Espaço), ou como é popularmente conhecida, NASA, foi fundada em 1958 nos Estados Unidos. A NASA é uma agência do Governo Federal dos Estados Unidos responsável pela pesquisa e desenvolvimento de tecnologias e programas de exploração espacial, sendo assim, a NASA tem como missão oficial a fomentação do futuro na pesquisa, descoberta e exploração espacial.

A fim de entender como o espaço funciona, a NASA é uma das principais agências atuante na área de pesquisa para efetuar a exploração desse ambiente desconhecido. Dessa maneira, missões são feitas periodicamente para tentar entender diversas indagações, as quais podem vir-a-ser desde como se desenvolve a criação de planetas ou estrelas, ou até em tentativas de descobrir novos seres vivos habitando tal ambiente.

Por decorrências de missões da agência diversas descobertas foram sendo concluídas, confirmando também teorias que ainda não haviam sido comprovadas cientificamente. No ano de 2020, a agência foi responsável por uma das descobertas mais intensas em relação à procura sobre seres vivos espaciais, a descoberta da molécula H2O no satélite natural do planeta Terra, a lua.[1].

Projetos de exploração e investigação por meio de satélites e sondas se tornaram rotineiros, como por exemplo a sonda Kepler, as sondas Surveyor e muitas outras. Porém, em 2012 o robô Curiosity, parte da missão Mars Science Laboratory (MSL), pousava no planeta vermelho, em Marte. Essa missão foi dada como uma das mais bem-sucedidas da história das agências, e também como uma das mais importantes até os dias de hoje.

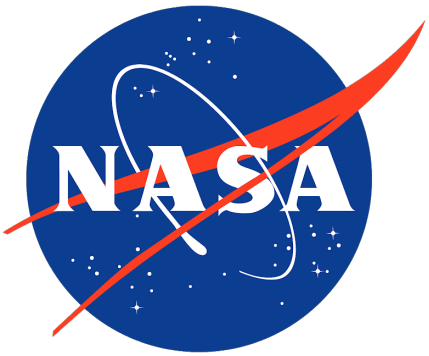

Figure 1: National Aeronautics and Space Administration

[1] Água presente na Lua: `moon.nasa.gov/news/155/theres-water-on-the-moon`

1.1 Marte

Marte é o quarto planeta do nosso Sistema Solar e o segundo menor, sendo vizinho próximo ao planeta Terra. Tal planeta é monitorado pela humanidade desde a antiguidade, onde os primeiros registros aconteceram por parte de astrônomos Egípcios[2], do qual também foi alvo de observações por parte de nomes como Aristóteles, Ptolomeu e Galileu Galilei.

O planeta vermelho, como é muitas vezes descrito, leva esse apelido pois o regolito, material que compõe a maior parte da superfície de um planeta, é composto por óxido de ferro, popularmente conhecido como ferrugem, gerando uma aparência avermelhada para o planeta. Dessa maneira, o óxido de ferro é formado quando o ferro puro é exposto ao oxigênio, entretanto ainda não existem respostas definitivas para explicar como ocorreu a formação de óxido de ferro em Marte.

Figure 2: Marte - 4º Planeta do Sistema Solar

A órbita de Marte encontra-se na conhecida Zona Habitável, da qual engloba a área entre Vênus até Marte. Essa zona pode ser considerável habitável pois é uma área da qual favorece o aparecimento de água no planeta. Entretanto, Marte carece de magnetosfera e possui uma atmosfera extremamente fina, portanto, o planeta possui pequena transferência de calor, sofre muito com ventos solares e pouca pressão atmosférica, tornando difícil a retenção da água em forma líquida no planeta. Logo, Marte pode ser considerado um planeta geologicamente morto.

No entanto, existem evidências de que o planeta tenha sido habitável no passado, mas o fato de que tenha albergado vida permanece incerto. Porém, com o decorrer das missões e envio de sondas e robôs, o estudo mais aprofundado sobre o ambiente do planeta vermelho pode se desenvolver de maneira mais rápida e eficiente.

Em julho de 2018, cientistas relataram a descoberta de um lago subglacial em Marte, o primeiro corpo estável de água conhecido no planeta. Ele fica a 1,5 km abaixo da superfície na base da calota polar sul e tem cerca de 20 quilômetros de largura. O lago foi descoberto usando o radar MARSIS a bordo da sonda Mars Express, e os dados foram coletados entre maio de 2012 e dezembro de 2015.[3] Dessa maneira, o estudo feito pelos robôs e sondas presentes em Marte ainda trazem esperanças sobre o fato de poder haver vida fora da Terra.

[2]Astrônomos Egípcios: https://ui.adsabs.harvard.edu/abs/2008POBeo..85...19N
[3]MARSIS: science.sciencemag.org/content/361/6401/490

1.2 Robô Curiosity

O robô Curiosity é um rover espacial, ou seja, um jipe robô, semelhante aos veículos Spirit e Opportunity. O Curiosity foi projetado para explorar a cratera Gale em Marte, como parte da missão Mars Science Laboratory (MSL) operada pela NASA. A sonda espacial foi lançada da Estação da Força Aérea de Cabo Canaveral em 26 de novembro de 2011, aterrissando na região Aeolis Palus, localizada dentro da **cratera Gale**, em Marte no dia 6 de agosto de 2012.

A história dessa tecnologia se iniciou em abril de 2004, quando a NASA solicitou à comunidade científica propostas de ideias de instrumentos científicos que pudessem ser instalados no Mars Science Laboratory (MSL é a designação de uma sonda espacial da NASA). Oito propostas foram selecionadas em 14 de dezembro daquele ano. Inicialmente, o lançamento estava previsto para 2009, porém a NASA decidiu adiar para 2011 sob a alegação de que faltavam alguns ajustes finais que dariam mais segurança à missão. Havia ainda uma discussão sobre a possibilidade de serem lançados dois ou três veículos idênticos para Marte.

Os objetivos da sonda incluem uma investigação do clima e da geologia marciana. O Curiosity transporta os mais avançados instrumentos científicos já utilizados em Marte, possibilitando à esta missão realizar análises do solo marciano nunca antes registradas. A comunidade internacional foi a responsável pelo fornecimento da maioria dos seus instrumentos, não tendo sido portanto um projeto exclusivo dos Estados Unidos.

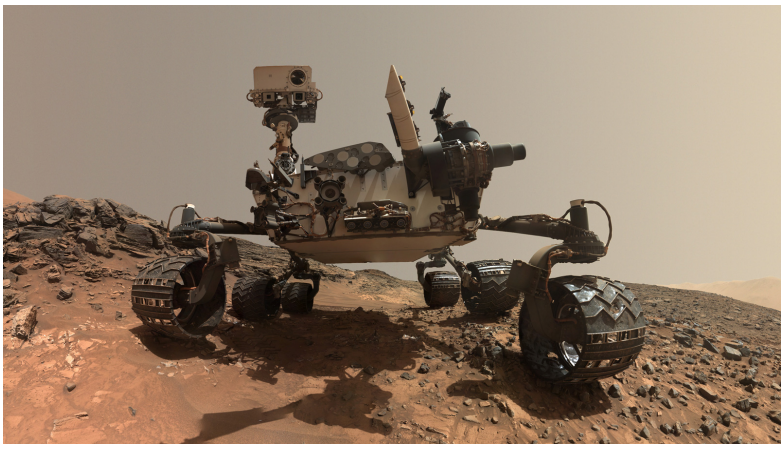

Figure 3: Rover Curiosity - MSL

O Curiosity possui velocidade máxima de 90km/h, massa de 899 quilogramas e altura de de 2,2 metros. Um fato interessante é que, em dezembro de 2012 a NASA decidiu que a missão do tal robô seria prorrogada indefinitivamente, e então serviria de base para o futuro rover, Mars 2020. Até então, o Curiosity foi capaz de fotografar quase 800 mil fotos, e percorreu quase 30 mil quilômetros e recentemente completou 3000 sols marcianos.

O robô possui diversas tecnologias acopladas a ele. Tais tecnologias têm como objetivo tentar coletar dados, investigar terrenos e substâncias e fazer testes, caso necessário. Portanto, o Curiosity pode ser considerado uma espécie de laboratório espacial em movimento, sendo que tais dados são enviados ao planeta Terra para que sejam analisados e sirvam de combustíveis para testes, previsões e conclusões de teorias sobre o planeta vermelho. O robô Curiosity apresenta diversos instrumentos científicos para, principalmente, coletar dados sobre Marte[4]. Os instrumentos são:

- **APXS:** Alpha Particle X-ray Spectrometer, ou em português, Espectrômetro de raios-x de partículas alfa. Esse instrumento tem como objetivo mensurar a abundância de elementos químicos em rochas ou do solo.

[4]Instrumentos do Curiosity: pds-geosciences.wustl.edu/missions/msl/index.htm

- **ChemCam:** Chemistry and Camera, ou em português, Química e Câmera. Esse instrumento tem como objetivo identificar a composição química e mineral de rochas e do solo por meio de um laser.
- **CheMin:** Chemistry and Mineralogy, ou em português, Química e Mineralogia. Esse instrumento tem como objetivo identificar e medir a abundância de vários minerais em Marte.
- **Dan:** Dynamic Albedo of Neutrons, ou em português, Albedo Dinâmico de Nêutrons. Esse instrumento tem como objetivo indicar a quantidade de água ligada ao solo ou rochas marcianas.
- **Engineering Cameras:** em português, Engenharia de Câmeras, são várias câmeras acopladas ao robô. O que pode ser considerado os olhos do rover.
- **MAHLI:** Mars Hand Lens Imager tem como objetivo revelar os minerais e texturas das superfícies das rochas por meio de lentes de aumento manual.
- **MARDI:** The Mars Descent Imager teve como objetivo gravar um vídeo colorido do terreno abaixo do Rover no momento de aterrissagem. O vídeo ajudou os planejadores da missão a selecionar o melhor caminho para o Curiosity quando o rover começou a explorar a cratera Gale.
- **MastCam**Mast Camera tem como objetivo fazer fotos e vídeos coloridos sobre o terreno marciano.
- **PLACES:** Esse instrumento tem como objetivo coletar dados de localização de mapas do rover.
- **RAD:** Radiation Assessment Detector, ou em português, Detector de avaliação de radiação, tem como objetivo medir o tipo e a quantidade de radiação prejudicial que atinge a superfície marciana do sol e de fontes espaciais.
- **REMS:** Rover Environmental Monitoring Station, ou em português, Estação de monitoramento Ambiental Rover, tem como objetivo coletar dados sobre o ambiente marciano.
- **SAM:** Sample Analysis at Mars, ou em português, Análise de Amostra em Marte, Esse instrumento tem como objetivo medir compostos químicos orgânicos e elementos leves que podem ser importantes ingredientes potencialmente associados à vida. Vale ressaltar que esse instrumento é composto por três diferentes instrumentos.
- **SPICE:** Spacecraft, Planet, Instrument, Pointing C-Matrix, and Event Kernels, ou em português, Nave espacial, Planeta, Instrumento, Apontando C-Matrix e Kernels de Eventos.

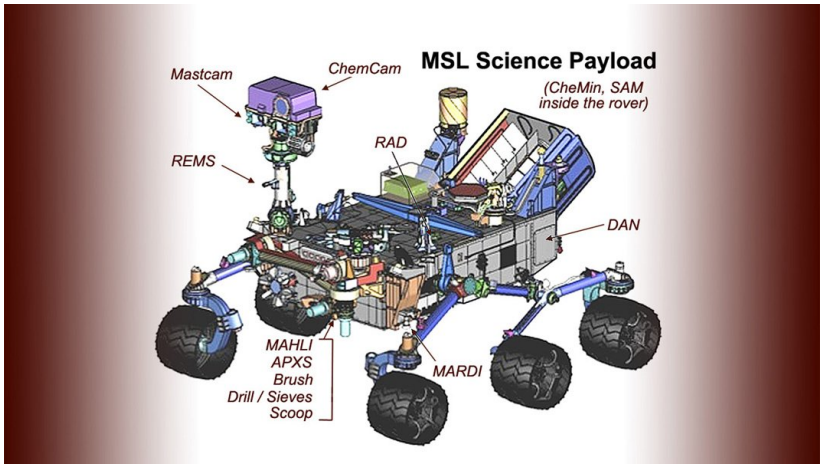

Figure 4: Rover Curiosity - Instrumentos

1.2.1 Estação de Monitoramento Ambiental Rover - REMS

Neste projeto foram utilizados os dados que são oriundos do instrumento REMS (Estação de Monitoramento Ambiental Rover)[5]. O REMS contém todos os instrumentos climáticos necessários para fornecer relatórios diários e sazonais sobre as condições meteorológicas ao redor do rover. Tais dados são importantes para entender variáveis climáticas e ambientais geradas pelo planeta vermelho.

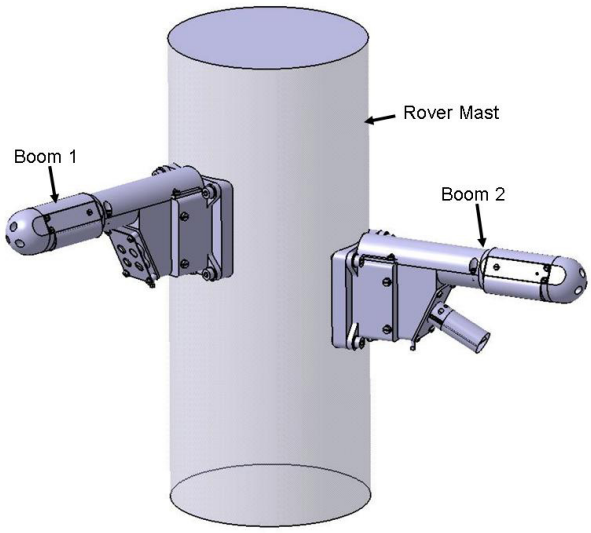

Figure 5: Estação de Monitoramento Ambiental Rover - REMS

O REMS pode ser considerado uma estação meteorológica com o objetivo de medir a pressão atmosférica, temperatura, umidade, ventos e níveis de radiação ultravioleta. Este instrumento foi projetado para sobreviver a uma faixa de temperatura de -130°C a +70°C e minimizar o consumo de energia para operação.

O instrumento REMS foi criado em parceria com o Centro de Astrobiologia (CAB) e o Conselho Superior de Investigações Nacional de Técnica Aeroespacial (CSIC- INTA), fornecido pelo governo Espanhol e então utilizado pelo rover marciano.

O REMS é dividido em dois pequenos cilindros que se estendem do mastro do rover. Como visto na figura 5, a barra 1 (Boom 1) contém diversos sensores infravermelhos que medem a intensidade da radiação infravermelha emitida pelo solo, fornecendo assim uma estimativa da temperatura do solo. Já a barra 2 (Boom 2) contém sensores de rastreamento da umidade atmosférica. Ambas as barras carregam sensores para medir a temperatura do ar.

Um sensor dentro do chassi do rover, exposto à atmosfera através de uma pequena "chaminé", mede as mudanças na pressão causadas por diferentes eventos meteorológicos. Um pequeno filtro protege o sensor contra contaminação por poeira. No convés do rover há uma série de detectores que são sensíveis a frequências específicas da luz solar. Estes medem a radiação ultravioleta na superfície marciana e as correlacionam com mudanças nas outras variáveis ambientais.

1.3 Conceitos Meteorológicos

A fim de introduzir o leitor a alguns conceitos que serão utilizados para chegar-se às análises efetuadas neste projeto, dedicou-se a criação deste tópico, o qual tem como objetivo explicar fatores de funcionamento na meteorologia, ambiente, solo e ar. Tais estudos foram feitos levando em consideração o ambiente terrestre, entretanto podem ajudar de alguma maneira a entender o campo meteorológico de Marte nas análises. Vale

[5]REMS Instrumentos - Clicável

ressaltar que tais inferências foram feitas utilizando os materiais dos professores Paulo Cesar Sentelhas e Luiz Roberto Angelocci [6] sobre estudos de meteorologia para os ambiente do solo e do ar.

1.3.1 Solo

De acordo com a Embrapa[7], no planeta Terra o solo é o resultado de um paciente trabalho da natureza. Partículas (minerais e orgânicas) vão sendo depositadas em camadas (horizontes) devido à ação do vento, do calor, do frio e comumente de organismos (fungos, bactérias, minhocas, formigas e cupins) que vão desgastando as rochas de forma lenta no relevo da terra. Os solos podem apresentar características e propriedades físicas, químicas e físico-químicas diferentes, podem ser argilosos, arenosos, vermelhos, amarelos ou cinza esbranquiçados. Eles ainda podem ser ricos ou pobres em matéria orgânica, e espessos (algumas dezenas de metros) ou rasos (alguns poucos centímetros), apresentando homogeneidade ou diferenças facilmente percebidas horizontalmente.

Porém, para a meteorologia, o regime térmico de um solo é determinado pelo aquecimento da superfície pela radiação solar e transporte, por condução, de calor sensível para seu interior. Durante o dia, a superfície se aquece, gerando um fluxo de calor para o interior. À noite, o resfriamento da superfície, por emissão de radiação terrestre, inverte o sentido do fluxo, que agora passa a ser do interior do solo para a superfície.

Fatores que são importantes na temperatura do solo são características das quais o fluxo de calor no solo dependem da condutividade térmica, do calor específico e da emissividade, os quais por sua vez dependem também do tipo do solo. Além disso, essa variação é afetada pela interação com outros fatores, dentre eles: Fatores Externos e Fatores Intrínsecos.

O tipo de solo também afeta a temperatura em tal ambiente, seja relacionado à textura, estrutura e teor de matéria orgânica do solo. Solos arenosos tendem a apresentar maiores amplitudes térmicas diárias nas camadas superficiais e menores em profundidade. Isso ocorre pelo fato dos solos arenosos terem maior porosidade, havendo um menor contato entre as partículas do solos, dificultando assim o processo de condução.

O relevo do solo é outro fator importante, pois este é um fator topoclimático, que condiciona o terreno a diferentes exposições à radiação solar direta e, também, ao acúmulo de ar frio durante o inverno. Em relação aos fatores climáticos, os terrenos de meia-encosta voltados para o norte (no hemisfério Sul) recebem mais energia do que os voltados para o sul. Já nas baixadas ocorre um maior acúmulo de ar frio durante o inverno, o que acaba condicionando redução da temperatura do solo também nessa área.

E finalmente, a variação térmica do solo pode estar associada à profundidade. Nas camadas mais superficiais, varia de acordo com a incidência de radiação solar. Em profundidades maiores, as máximas tendem a ocorrer mais tarde, assim como as mínimas.

[6] Temperatura do ar e do solo
[7] Embrapa

1.3.2 Ar

O ar, para o planeta Terra, é a mistura gasosa que forma a atmosfera terrestre. Para além do vapor de água que aparece em diferentes proporções, este fluido é composto por 78 partes de nitrogênio, 21 partes de oxigênio e 1 de argão e outros gases similares, bem como umas centésimas de dióxido de carbono.

A temperatura do ar é um dos efeitos mais importantes da radiação solar. O aquecimento da atmosfera próxima à superfície terrestre ocorre principalmente por transporte de calor, a partir do aquecimento da superfície pelos raios solares. O transporte de calor sensível (H) na atmosfera dá-se por 2 processos: Condução Molecular e Difusão Turbulenta.

A condução molecular é um processo lento de troca de hidrogênio (H), ocorrendo pelo contato entre as moléculas de ar. Assim, esse processo tem extensão espacial limitada, ficando restrito à camada limite superficial. Já a difusão turbulenta é um processo rápido de troca de energia, em que parcelas de ar aquecidas pela superfície entram em movimento convectivo desordenado, transportando calor (H), vapor (LE), e entre outros, para camadas superiores da atmosfera.

Fatores que são importantes na temperatura do ar são aqueles associados às três escalas dos fenômenos atmosféricos, ou seja: Fatores Macroclimáticos, Fatores Topoclimáticos e Fatores Microclimáticos. A temperatura do ar varia basicamente em função da disponibilidade de radiação solar na superfície terrestre. O valor máximo diário da temperatura do ar ocorre normalmente de 2 a 3h após o pico de energia radiante, o que se deve ao fato da temperatura do ar ser medida a cerca de 1,5 a 2,0 m acima da superfície. Já a temperatura mínima diária ocorre de madrugada, alguns instantes antes do nascer do sol.

Deve-se levar em consideração também a variabilidade espacial (horizontal), que é basicamente definida pelos fatores determinantes do clima, como latitude, altitude, continentalidade, correntes oceânicas, massas de ar e outros fatores. A temperatura do ar varia espacialmente também na vertical. Tanto o aquecimento quanto o resfriamento do ar se dão a partir da superfície, logo, durante o dia a tendência é da temperatura do ar ser maior próxima à superfície, e menor com a altura. Já de madrugada, essa situação inverte-se, sendo a temperatura menor próxima à superfície e maior com o aumento da altura.

2 CRISP-DM

Com o passar dos anos o avanço tecnológico foi eminente, tal ascensão fez com o que o volume dos dados crescesse de maneira astronômica. Sendo assim, armazenar tais dados se tornou uma tarefa vital. Com isso, a mineração de dados entra como principal atuante para extrair informações dos dados armazenados, aplicando algoritmos ou ferramenta em tal ambiente. Entretanto, como em todo processo, a mineração de dados precisa solucionar problemas por meio de diagnósticos, análises e planejamento. Dessa maneira, esse projeto irá utilizar todas as etapas presente na metodologia CRISP-DM referente a cada um dos seis tópicos para a confecção deste projeto.

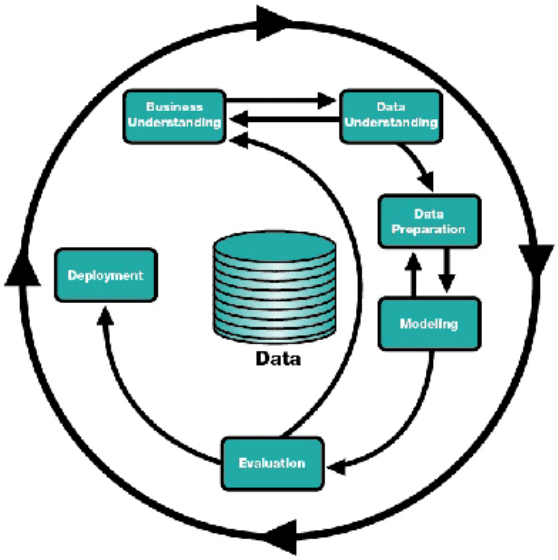

Figure 6: CRISP-DM consortium - 2000

- **Business Understanding (Entendimento do Negócio):** esta fase se concentra no entendimento dos objetivos e requisitos do projeto sob uma perspectiva de negócios, convertendo esse conhecimento em uma definição de problema de mineração de dados e em um plano preliminar para atingir os objetivos.
- **Data Understanding (Compreensão dos Dados):** esta fase consiste em organizar e documentar todos os dados que se encontram disponíveis.
- **Data Preparation (Preparação dos Dados):** nesta fase com os dados já identificados, documentados e analisados, é hora de aplicar a parte técnica de análise.
- **Modeling (Modelagem):** é nesta fase que são aplicadas de fato as técnicas de Data Mining, com base nos objetivos identificados no primeiro momento.
- **Evaluation (Avaliação):** nesta fase já teremos um modelo ou modelos a partir de uma análise perspectiva dos dados. Antes de prosseguir para a implantação final do modelo, realizaremos uma avaliação completa e revisaremos as etapas executadas para criá-lo, garantindo que o modelo atinja adequadamente aos objetivos de negócios.
- **Deployment (Implementação do Modelo):** aqui teremos a criação do modelo, mas geralmente não é o fim do projeto. Mesmo que o objetivo do modelo seja aumentar o conhecimento dos dados, o conhecimento adquirido precisará ser organizado e apresentado de forma que o cliente possa usá-lo. Isso geralmente envolve a aplicação de modelos "ativos" nos processos de tomada de decisão de uma organização.

3 Entendimento do Negócio

A fim de entender como o ambiente e o clima se desenvolveram em Marte nos mais de 9 anos de exploração concluídos pelo rover Curiosity, dedicou-se perguntas norteadora das quais serão utilizadas para guiar todo o processo de criação desse projeto, bem como guiará as etapas do CRISP-DM. Portanto, as perguntas norteadoras serão:

- De acordo com os dados encontrados pela API do rover, quais características climáticas Marte apresentou nos anos de exploração?
- Dadas as variáveis, é possível fazer uma previsão em relação à temperatura média máxima do solo de Marte?

Para a conclusão justa de tais indagações, deve-se levar em consideração as métricas para o sucesso de tais respostas. Os entregáveis para a conclusão da primeira pergunta será o desenvolvimento de uma análise exploratória sobre os dados disponíveis, e então a formulação de uma relatório resumido, do qual tem como objetivo transcrever a exploração feita pelo rover em todos esses anos de exploração, contendo informações sobre o clima, dos quais foram gerados pelos dados presentes na API do Curiosity. Portanto, de maneira lógica, esse relatório deve ser finalizado ao fim da análise exploratória.

Os entregáveis em relação à segunda pergunta concluí-se em um modelo preditivo, por meio da utilização de técnicas de regressão. O modelo em questão, ao final, deve conseguir prever a temperatura máxima do solo em Marte. Tal objetivo foi traçado pois o solo marciano é o ambiente que consegue melhor conter o calor (o que será provado na etapa da exploração dos dados). Portanto, será interessante saber quais variáveis numéricas influenciam na temperatura máxima do solo marciano, e dado alguns valores, prever qual será tal temperatura. As métricas de sucesso serão avaliadas pelas métricas de regressão, Desvio Médio Quadrático (RMSE) e o Erro Médio Absoluto (MAE), uma vez que como será apresentado na etapa de avaliação, o Erro Médio Quadrático (MSE) não seria a melhor métrica para avaliar a regressão aplicada nesse projeto. Portanto, tais métricas de regressão RMSE e MAE deverão estar no intervalo entre 0.3 e 5, isso quer dizer que, a regressão para prever a temperatura máxima do solo pode errar em 0.3 até 5 graus Celsius.

Na etapa de entendimento do negócio, é válido ressaltar as ferramentas das quais serão utilizadas para a confecção e conclusão desse projeto. Dessa maneira, a parte majoritária desse projeto usará programação em linguagem Python em sua versão 3.8, em conjunto com as respectivas bibliotecas:

- **json:** Será utilizado para coletar os dados em formato JSON na API.
- **requests:** Será utilizado para acessar os dados na API.
- **pandas:** Será utilizado para a manipulação dos dados.
- **seaborn:** Será utilizado para criar interfaces gráficas sobre os dados.
- **matplotlib:** Será utilizado para criar interfaces gráficas sobre os dados.
- **sqlalchemy:** Será utilizado para alocar, e posteriormente, acessar os dados no banco de dados.
- **scikit-learn:** Será utilizado para a criação do modelo de regressão linear.

Dessa maneira, serão utilizados os bancos de dados PostegreSQL e SQLite para alocar os dados referentes à API de maneira única, dos quais serão coletados por meio da linguagem Python. E finalmente, a documentação do projeto será efetuada por meio do software LaTeXpara uma melhor organização e estruturação do projeto. Destaca-se que o código usado para a criação deste trabalho estará anexado ao fim desse documento. Após concluída as perguntas norteadoras na etapa do entendimento do negócio, será possível a criação de análises mais focadas e consistentes, levando em consideração um único objetivo final, do qual será capaz de responder as indagações feitas anteriormente por meio de técnicas claras e responsáveis sobre os dados disponíveis para tal.

4 Entendimento dos dados

Antes de entender os dados em si, é necessário documentar como chegou-se à captura deles. Os dados utilizados neste projeto foram extraídos por meio de uma API que pode ser encontrada acessando o site MSL [8] avaliando a lista de todos os requerimentos que a página está fazendo (Network), por meio da ferramenta de desenvolvedor presente no navegador em questão.

Acessando a API do qual o site faz o requerimento dos dados, é possível evidenciar que tais dados estão em formato JSON, um formato comum quando se trata de API's. Porém, antes da captura dos dados, percebe-se um aviso de domínio por parte do Centro de Astrobiologia (CAB), do qual possui a responsabilidade dos dados gerados pelo instrumento REMS, presente no Curiosity. Destaca-se que o período da coleta dos dados se deu, desde o inicio em 07 de Agosto de 2012 até 07 de Abril de 2021. Sendo assim, houveram 3082 sois desde a chega do rover em solo marciano.

4.1 Dicionário dos dados

O dicionário de dados visa explicar as variáveis presentes no banco de dados abordado neste projeto. Dessa maneira, para facilitar o trabalho de exploração de dados, cria-se um dicionário de dados, do qual tem como objetivo deixar as informações mais claras para o leitor sobre o que significa determinada variável. Portanto, a base de dados detêm as seguintes variáveis[9]:

- **ID:** corresponde ao ID do conjunto dos dados gerados em determinado Sol.
- **Terrestrial Date:** data terrestre.
- **LS:** Longitude Solar. A longitude solar é um ângulo que fornece a posição de Marte em sua órbita.
- **Sol:** como o dia em marte possui mais de 24h, não é possível igualá-lo ao dia terrestre. Vale ressaltar que, por conta disso, o ano em marte possui 687 dias. Sendo assim, a nomenclatura Sol equivale à data marciana.
- **Season:** considerado o mês marciano. Os meses marcianos variam de 46 a 67 sóis (dias marcianos) de duração.
- **Min Temp:** temperatura mínima do ar marciano (em graus Celsius).
- **Max Temp:** temperatura máxima do ar marciano (em graus Celsius).
- **Pressure:** pressão atmosférica marciana.
- **Pressure String:** grau categórico da pressão.
- **Abs Humidity:** umidade do ar marciano.
- **Wind Speed:** velocidade do vento em Marte.
- **Wind Direction:** direção do vento em Marte.
- **Atmo Opacity:** grau categórico de opacidade atmosférica, podendo ser:
  - Dust Devils and Strong Winds: ventos fortes ocasionados por tempestade de areia.
  - Fog: névoa.
  - Frost: geada.
  - Ice and Fog: neve com névoa.
  - Snow: neve.

[8] MSL: mars.nasa.gov/msl/weather/
[9] CAB: cab.inta-csic.es/rems/rems_weather.xml

  - Storm: tempestade.
  - Sunny and Cloudy: ensolarado com nuvens.
  - Sunny: ensolarado.
  - Windy: ventoso.

- **Sunrise:** nascer do sol em Marte.
- **Sunset:** pôr do sol em Marte.
- **Local UV Irradiance Index:** indicador da intensidade da radiação ultravioleta do Sol no local onde o rover está presente.
- **Min Gts Temp:** temperatura mínima do solo marciano (em graus Celsius).
- **Max Gts Temp:** temperatura máxima do solo marciano (em graus Celsius).

Após todas as variáveis terem sido informadas sobre sua natureza, é válida a explicação mais aprofundada sobre alguma dessas para que o leitor entenda mais detalhadamente, também servindo como uma exposição de possíveis curiosidades, principalmente ao que diz sobre medidas.

- **Sol:** um sol (marciano) equivale a cerca de 24 horas e 40 minutos. Para o rover Curiosity, sol 0 corresponde ao dia de sua chegada em Marte.
- **Temperaturas:** Marte está mais longe do Sol do que a Terra, isso faz com que Marte seja mais frio que o nosso planeta. Além disso, a atmosfera marciana, que é extremamente tênue, não retém o calor; portanto, a diferença entre as temperaturas diurnas e noturnas varia mais do que em nosso planeta.
- **Pressão:** a pressão é uma medida da massa total em uma coluna de ar acima de nós. Como a atmosfera de Marte é extremamente tênue, a pressão na superfície de Marte é cerca de 160 vezes menor do que a pressão na Terra. A pressão média na superfície marciana é de cerca de 700 Pascais (100.000 Pascais na Terra).
- **Radiação UV:** o índice de irradiância ultravioleta local (UV) é um indicador da intensidade da radiação ultravioleta do Sol no local Curiosity. A radiação ultravioleta é um agente prejudicial para a vida. Na Terra, a camada de ozônio impede que a luz ultravioleta prejudicial alcance a superfície, para o benefício de plantas e animais. No entanto, em Marte, devido à ausência de ozônio na atmosfera, a radiação ultravioleta atinge a superfície marciana.
- **Longitude Solar:** um ano marciano dura cerca de dois anos da Terra, que é o tempo que Marte leva para orbitar o sol. Então, a longitude solar é um ângulo que fornece a posição de Marte em sua órbita.
- **Nascer e Pôr do Sol:** a duração de um dia marciano (Sol) é de cerca de 24 horas e 40 minutos. A duração da luz do dia varia ao longo do ano marciano, assim como na Terra.

Após entendido o caminho até os dados, dicionário dos dados concluído e demais explicações ocasionais definidas, encerra-se a etapa conceitual do entendimento dos dados. Portanto, será dada continuidade em busca das respostas das perguntas de negócios.

4.2 Análise Exploratória

Nesse momento, será feita a análise exploratória dos dados, a fim de entendê-los principalmente numericamente, ou seja, como estão associados e extrair possíveis informações para responder às perguntas de negócios definidas anteriormente. É válido ressaltar que toda a análise exploratória foi efetuada utilizando as bibliotecas Pandas e Seaborn na linguagem Python.

4.2.1 Manipulação dos tipos de dados

Inicialmente, como era de esperar, as variáveis foram interpretadas pelo Python como do tipo objeto. Sendo assim, nesse estágio inicial da análise exploratória, tornou-se necessário efetuar a troca dos tipos de dados das variáveis que não correspondem ao tipo mais adequado. Após a mudança de tipagem dos dados, conclui-se o tipo das variáveis da seguinte maneira:

```
RangeIndex: 2937 entries, 0 to 2936
Data columns (total 8 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   DataTerra    2937 non-null   datetime64[ns]
 1   DataMarte    2937 non-null   Int64
 2   MaxTempAr    2937 non-null   Int64
 3   MaxTempSolo  2937 non-null   Int64
 4   MinTempAr    2937 non-null   Int64
 5   MinTempSolo  2937 non-null   Int64
 6   Pressao      2937 non-null   Int64
 7   NivelUV      2937 non-null   object
dtypes: Int64(6), datetime64[ns](1), object(1)
```

Figure 7: Tipos das variáveis

Conforme explicado na seção 4.1, a variável DataMarte é referente à data do planeta Marte. Sendo assim, faz sentido transformá-la para o tipo inteiro, uma vez que, na nomenclatura oficial, "Sol 1" significa dia 1 (desde o pouso do rover em solo marciano até então). Vale frisar que tal maneira de contar os dias em Marte desenvolveu-se assim, pois o dia em Marte possui mais de 24 horas, portanto, não seria possível utilizar a data da maneira terrestre para Marte.

Uma hipótese que poderia ser levantada é a de que, na etapa de manipulação dos dados, resumisse os dados das temperaturas máximas do solo e ar em uma só, e posteriormente resumir as temperaturas mínimas do solo e ar, concatenando apenas duas novas variáveis para temperatura. Entretanto, dado que Marte possui atmosfera tênue, a utilização de tal artifício não seria fiel às análises, uma vez que, por possuir uma atmosfera nesse estado, as temperaturas do ar e do solo podem ser diferentes.

Com os dados manipulados para seus devidos tipos, as análises farão mais sentido e serão mais concisas, dado que a base possuía diversos dados do tipo inteiro, dos quais inicialmente foram compreendidos como do tipo objeto. Portanto, após esta breve manipulação, pode-se seguir para análises estatísticas, utilizando métodos de estatística descritiva sobre os dados, gráficos e afins, para que então o arranjo dos dados sirva de resposta às perguntas de negócio.

4.2.2 Estatística Descritiva

Nesse momento serão efetuadas as análises que envolvem métodos estatísticos para a validação. Inicia-se com a análise de medidas de tendência central e de dispersão, do qual tem o objetivo de representar o centro da base de dados e como estes estão dispersos entre si. A medida de tendência central é composta principalmente pela média, mediana, moda e os percentis. Entretanto, é de se levar em consideração que algumas medidas de tendência central sofrem com valores discrepantes em relação aos outros, denominados "outliers", nesse caso por exemplo, a média é afetada por tal. Portanto, a seguir estão as medidas de tendência central da base de dados referentes às variáveis numéricas.

	DataMarte	MaxTempAr	MaxTempSolo	MinTempAr	MinTempSolo	Pressao
count	2937.000000	2937.000000	2937.000000	2937.000000	2937.000000	2937.000000
mean	1552.811713	-12.505958	2.779367	-74.854273	-79.834866	831.365339
std	885.258299	10.415355	9.081808	5.665086	8.946269	56.093450
min	1.000000	-54.000000	-61.000000	-100.000000	-136.000000	707.000000
25%	793.000000	-22.000000	-4.000000	-79.000000	-85.000000	791.000000
50%	1550.000000	-11.000000	4.000000	-74.000000	-79.000000	845.000000
75%	2319.000000	-4.000000	10.000000	-71.000000	-74.000000	875.000000
max	3082.000000	11.000000	24.000000	-52.000000	-8.000000	925.000000

Figure 8: Medidas de tendência central e dispersão

Destaca-se que o rover coletou dados de temperatura do ar e do solo. Sendo assim, no ambiente proveniente em Marte, tais temperaturas obviamente possuem valores distintos, pois, como esclarecido, o planeta vermelho possui atmosfera amena. Portanto, os ambientes do ar e do solo não conseguem reter o calor tão bem, fazendo-as variar bastante, ao comparar-se à Terra.

É possível verificar na figura 8 que, dos 3082 sois presenciados pelo rover em solo marciano, foi possível recolher dados de apenas 2937 sois. Sendo assim, em todo o período da coleta dos dados, de 12 de Agosto de 2012 até 07 de Abril de 2021, houveram 145 dias dos quais o rover não retornou dados à API.

4.2.3 Temperaturas Máximas

Em relação à **Temperatura Máxima do Ar** em Marte, percebe-se que houveram 2937 registros coletados. A média da temperatura máxima do ar foi igual a -12ºC. Diante dos dados, é possível concluir que, em certos dias marcianos, a temperatura máxima do ar mais baixa registrada até então foi igual a -54ºC. Entretanto, o dia mais quente em relação à temperatura do ar registrou uma temperatura de 11ºC. Da mesma maneira, é possível concluir que, dos 2937 registros em relação à temperatura do ar, 1468 dias, ou seja, a mediana, obteve temperaturas menores ou iguais a -11ºC.

Em relação à **Temperatura Máxima do Solo** em Marte, percebe-se que houveram 2937 registros coletados. A média da temperatura máxima do solo foi igual a 2ºC. Diante dos dados, é possível concluir que, em certo dia marciano, a temperatura máxima do solo mais baixa registrada até então foi igual a -61ºC, dado estranho uma vez que o solo retém calor melhor, um possível outlier. Entretanto, o dia mais quente em relação à temperatura do solo registrou uma temperatura de 24ºC. Da mesma maneira, é possível concluir que, dos 2937 registros em relação à temperatura do solo, 1468, ou seja, a mediana, obteve temperaturas menores ou iguais a 4ºC.

Então, até o momento é necessário investigar a temperatura encontrada para o ar e solo das quais foram iguais a -54ºC e -61ºC respectivamente, pois tal evento assemelha-se a um possível outlier, da qual pode ser visto melhor em gráficos de linha.

4.2.4 Temperaturas Máximas - Gráficos

A fim de entender e visualizar a distribuição dos dados e das frequências das temperaturas máximas do ar e solo, utilizou-se gráficos como de linha e histograma, do qual tem como objetivo tornar a visualização das análises mais esclarecedora em relação aos dados dos quais foram analisado anteriormente para temperaturas máximas no planeta marciano. A seguir, gerou-se gráficos de linhas com as temperaturas máximas do ar e do solo coletados nos dias marcianos (a data é nomeada como Sol).

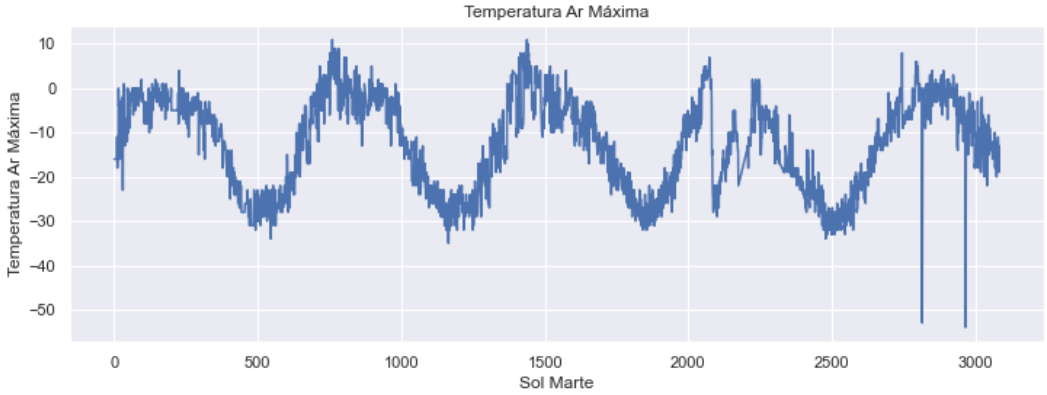

Figure 9: Temperatura Máxima Ar por Sol Marciano

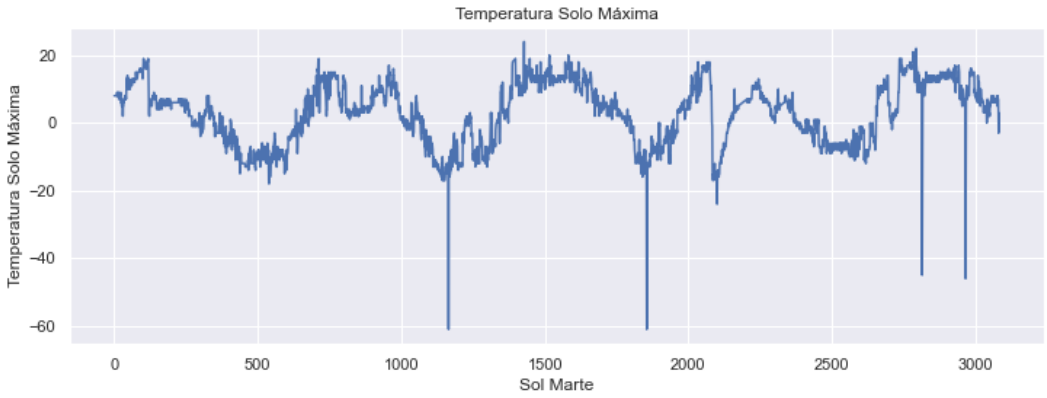

Figure 10: Temperatura Máxima Solo por Sol Marciano

Portanto, percebe-se que as temperaturas máximas do solo possuem alguns outliers, algo extremamente estranho uma vez que houve uma queda brusca em relação ao histórico dos dados (o que talvez possa ser considerado uma falha na coleta dos dados). Dessa maneira, como tinha-se pensado anteriormente na etapa de análise das estatísticas descritivas, a temperatura do ar realmente possui temperaturas máximas menores em relação às temperaturas máximas do solo, uma vez que o mínimo da temperatura máxima do solo era na verdade outlier, como analisado na temperatura igual a -61^{o}C para o solo. Dessa forma, tais outliers deverão ser tratados posteriormente na parte de preparação dos dados.

Então, conclui-se após analisar os gráficos de linha das temperaturas máximas em Marte, que, temperaturas mais quentes ocorreram com mais frequência de dados coletados do solo. Ou seja, aparentemente o solo se mostra um ambiente melhor de absorção de calor do que em relação ao ar. No entanto, dado que o ano em Marte possui 687 dias, percebe-se que há uma certa repetição da distribuição dos dados a cada ano em relação às temperaturas máximas do ar e do solo.

Verificando os histogramas das variáveis de temperatura máxima do ar e do solo, pode-se tornar esclarecedor as temperaturas máximas mais frequentes nos dados coletados pelo rover no planeta marciano, dado o período em questão. Dessa maneira, os histogramas se desenvolveram da seguinte maneira.

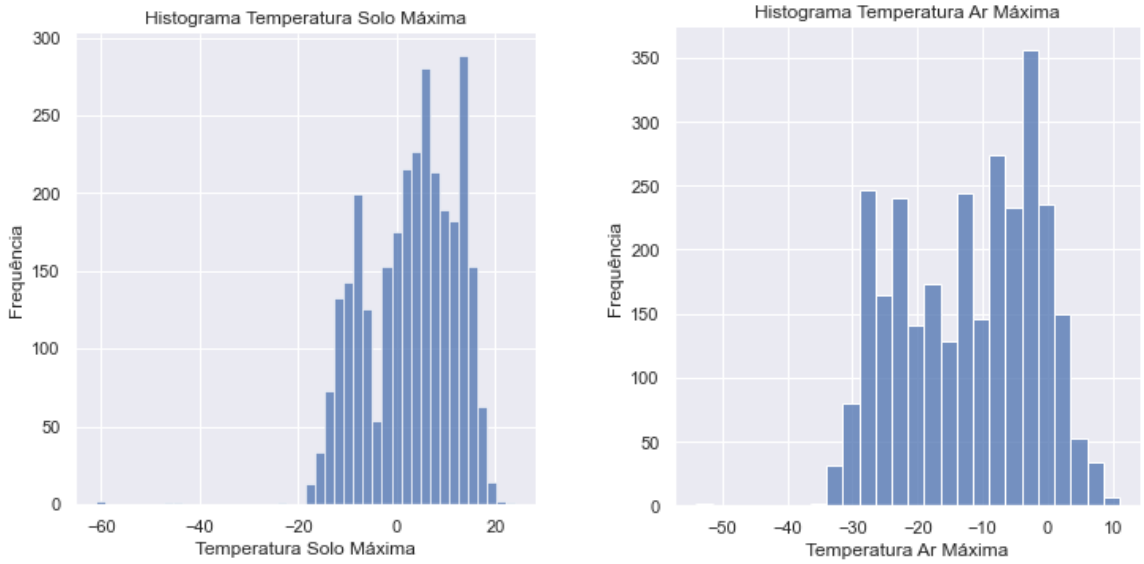

Figure 11: Histogramas Temperaturas Máximas

Então, conclui-se após analisar os histogramas para as variáveis de temperatura máxima do solo e ar, que, em relação a temperaturas máximas do solo, as mais frequentes estavam no intervalo entre -20ºC a 20ºC, porém mais frequentes no intervalo de 0ºC a 20ºC, e percebe-se alguns outliers, como por exemplo, certo dia que a temperatura máxima no solo foi de -61ºC, como evidenciado na figura 9. Já em relação às temperaturas máximas do ar, tal ambiente obteve temperaturas mais frequentes no intervalo entre -30ºC a 10ºC, demonstrando que o ar possui temperaturas mais variadas, ou seja, sem uma concentração perceptível, dada a falta de atmosfera no planeta vermelho.

4.2.5 Temperaturas Mínimas

Em relação à **Temperatura Mínima do Ar** em Marte, percebe-se que houveram 2937 registros coletados. A média da temperatura máxima do ar foi igual a -74ºC. Diante dos dados, é possível concluir que, em certo dia marciano, a temperatura máxima do ar mais baixa registrada até então foi igual a -100ºC. Entretanto, o dia com maior temperatura mínima em relação ao ar registrou uma temperatura de -52ºC. Da mesma maneira, é possível concluir que, dos 2937 registros em relação à temperatura do ar, 1468 dias, ou seja, a mediana, obteve temperaturas menores ou iguais a -71ºC.

Em relação à **Temperatura Mínima do Solo** em Marte, percebe-se que houveram 2937 registros coletados. A média da temperatura máxima do solo foi igual a -79ºC. Diante dos dados, é possível concluir que, em certo dia marciano, a temperatura máxima do solo mais baixa registrada até então foi igual a -136ºC. Entretanto, o dia com maior temperatura mínima em relação ao solo registrou uma temperatura de -8ºC. Da mesma maneira, é possível concluir que, dos 2937 registros em relação à temperatura do solo, 1468 ,ou seja, a mediana, obteve temperaturas menores ou iguais a -74ºC.

Então, depreende-se que, como explicado anteriormente, como Marte possui atmosfera rasa, os dias possuem grande amplitude térmica, fazendo muito frio à noite quando o Sol se põe. Dessa maneira, evidencia-se que, em relação às temperaturas mínimas, o solo desenvolve temperaturas menores do que em relação ao ar pela capacidade de absorção de temperatura do qual tal ambiente possui, juntamente com a falta de atmosfera resultando na grande amplitude térmica.

4.2.6 Temperaturas Mínimas - Gráficos

A fim de entender e visualizar a distribuição dos dados e das frequências das temperaturas mínimas do ar e solo, utilizou-se gráficos como de linha e histograma, do qual tem como objetivo tornar a visualização das análises mais esclarecedora em relação aos dados dos quais foram analisados anteriormente para temperaturas mínimas no planeta marciano. A seguir, gerou-se gráficos de linhas com as temperaturas mínimas do ar e do solo coletados nos dias marcianos.

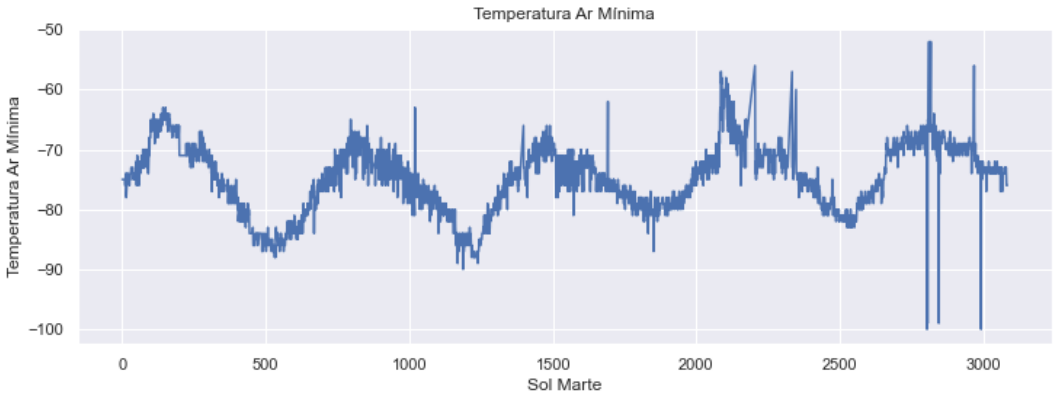

Figure 12: Temperatura Mínima Ar por Sol Marciano

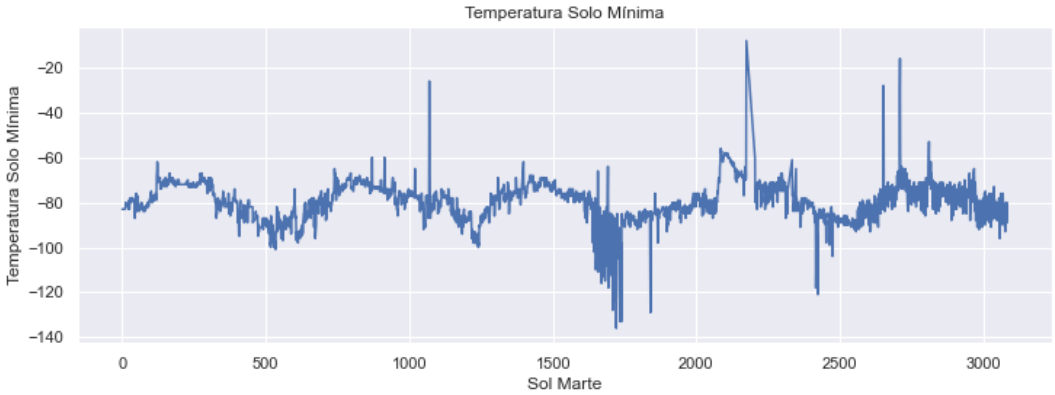

Figure 13: Temperatura Mínima Solo por Sol Marciano

Portanto, percebe-se que as temperaturas mínimas do ar possuem alguns outliers, algo extremamente estranho uma vez que houve uma queda brusca em relação à temperatura mínima (o que talvez possa ser considerado uma falha na coleta dos dados). Tais outliers afetaram tanto a máxima e mínima, retornando dados enganosos para tal ambiente. Percebe-se também, no gráfico do solo, na figura 13, no período do sol 1500 ao sol 200, quedas constantes em relação à temperatura mínima do solo. Dessa maneira, como tinha-se concluído anteriormente na etapa de análise das estatísticas descritivas, a temperatura do solo possui temperaturas mínimas menores em relação às temperaturas mínimas do ar.

Então, conclui-se após analisar os gráficos de linha das temperaturas mínimas em Marte, que, temperaturas mais frias ocorreram com mais frequência nos dados coletados do solo. Algo que faz sentido, dado o fato que Marte é um planeta arenoso e desértico, portanto o solo possui amplitude térmica alta. No entanto, dado que o ano em Marte possui 687 dias, percebe-se que há uma certa repetição da distribuição dos dados a cada ano em relação às temperaturas mínimas do ar.

Verificando os histogramas das variáveis de temperatura máxima do ar e do solo, pode-se tornar esclarecedor as temperaturas máximas mais frequentes nos dados coletados pelo rover no planeta marciano, dado o período em questão. Dessa maneira, os histogramas se desenvolveram da seguinte maneira.

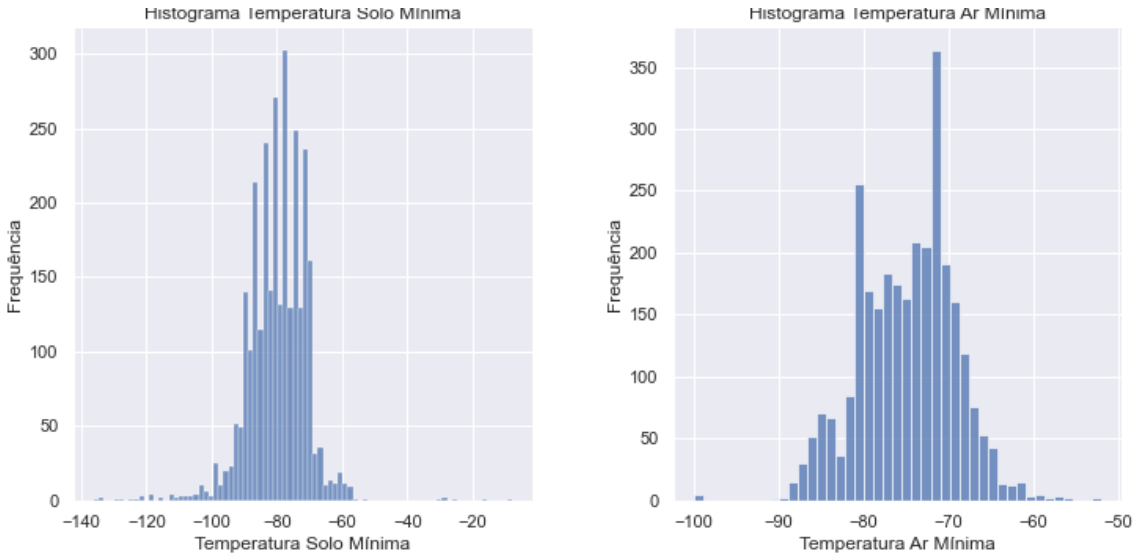

Figure 14: Histogramas Temperaturas Mínimas

Na figura 14 verifica-se os histogramas para as variáveis temperaturas mínimas do ar e solo coletados pelo rover Curiotisty até 07 de Abril. Inicialmente, é possível inferir que a temperatura mínima do solo possuiu temperaturas com frequências mais variadas do que em relação às temperaturas mínimas do ar. Entretanto, as frequências das temperaturas mínimas do solo variam, principalmente, no intervalo entre -100ºC a -70ºC. Em relação às frequências da temperatura mínima do ar, há variação no intervalo entre -90ºC a -60ºC.

Consequentemente, é possível verificar que as variáveis de temperatura mínima possuem muito mais outliers do que as temperaturas máximas coletadas pelo rover em Marte. Isso pode significar que, em Marte, temperaturas mínimas podem alcançar valores extremos, entretanto ainda deve-se levar em consideração os intervalos explicados anteriormente, dos quais são as temperaturas mais frequentes de acordo com a coleta dos dados.

Portanto, em relação a todos os dias em que o rover Curiosity coletou os dados de temperaturas mínimas, o mesmo presenciou mais frequentemente temperaturas do solo entre -100ºC a -70ºC. Já em relação às temperaturas mínimas do ar, o rover presenciou de maneira mais frequente temperaturas entre -90ºC a -60ºC.

4.2.7 Temperaturas - Conclusão

Então, visualizando novamente as figuras 9, 10, 13, 12, percebe-se, de maneira majoritária, porém podendo haver alguns históricos de dados incondizentes ao respectivo ciclo, ciclos em relação às temperaturas, sejam máximas ou mínimas e nos ambientes de solo ou ar. Interessante ressaltar que, tais ciclos são se repetem à cada, aproximadamente, 680 dias, ou seja, o que configura um ano em Marte.

Importante ressaltar a detecção de dados dos quais não são condizentes aos respectivos ciclos captados a cada ano. Dessa maneira pode-se levantar algumas possibilidades, a primeira em relação à falha na coleta dos dados em relação à API, configurando uma confusão em relação ao histórico dos dados. A segunda possibilidade pode ser referente às viagens feitas pelo rover no planeta Marte, da qual o Curiosity pode ter chegado a determinado local na cratera de Gale, onde a missão do rover está limitada, do qual captou-se

temperaturas distintas ao usual histórico dos dados coletados anteriormente.

Portanto, após concluir o entendimento e visualização dos gráficos de temperatura, pode-se concluir inferencialmente que, em relação à cratera de Gale, as temperaturas do solo e do ar realmente são diferentes entre si, cabendo às temperaturas referentes ao solo absorção melhor em relação ao calor, entretanto, pelo mesmo efeito que ocorre em ambientes desérticos, quando o sol se põe, o solo desenvolve temperaturas mais frias. Dessa forma, temperaturas máximas maiores ocorrem no ambiente do solo, e temperaturas mínimas menores ocorrem no ambiente do solo também. E finalmente, tornou-se possível a constatação dos ciclos das temperaturas a cada ano em Marte, bem como possíveis falhas na coleta dos dados (referentes aos outliers).

4.2.8 Pressão

A pressão atmosférica em Marte é um fator climático muito relevante. De maneira simples, a pressão atmosférica é o peso que o ar exerce sobre a superfície terrestre. Sua manifestação está diretamente relacionada à força da gravidade e à influência que essa realiza sobre as moléculas gasosas que compõem a atmosfera. No planeta Terra, a pressão atmosférica muitas vezes serve como norteadora para determinado acontecimento climático. Sendo assim, a existência da pressão atmosférica, e a variação de seus valores entre as diferentes áreas da superfície terrestre, são características que influenciam diretamente a dinâmica climática. Basicamente, ela interfere em algumas condições meteorológicas básicas, como os ventos, as temperaturas e a precipitação.

Entretanto, como foi enfatizado nesse projeto, o planeta marciano possui atmosfera extremamente tênue. Portanto, a pressão na superficie de Marte é cerca de 160 vezes menor em comparação a pressão na Terra. A pressão media na superfície marciana (coletada por outros aparelhos), de acordo com o CAB, é de cerca de 700 Pascais, diferente da média encontrada nesse projeto pois o rover Curiosity encontra-se majoritariamente na cratera de Gale, que é mais funda.

Dessa forma, em relação aos dados coletados referentes a pressão atmosférica marciana, dos quais foram coletados pelo rover majoritariamente na cratera de Galera, pode-se concluir que houve 2937 registros. A média da pressão atmosférica foi igual a 831 pascais. Diante dos dados, é possível concluir que, em certo dia marciano, a pressão máxima na cratera Galera foi de 925 pascais, já a mínima foi equivalente a 707 pascais. Pode-se concluir também que, na cratera Gale 1468 dias marcianos obtiveram pressão atmosférica igual ou menor a 845 pascais, ou seja, a mediana da pressão.

Consequentemente, explorando os dados coletados da API do rover Curiosity, os dados das pressões diárias, em sol marciano, se desenvolveram da seguinte maneira no período aplicado:

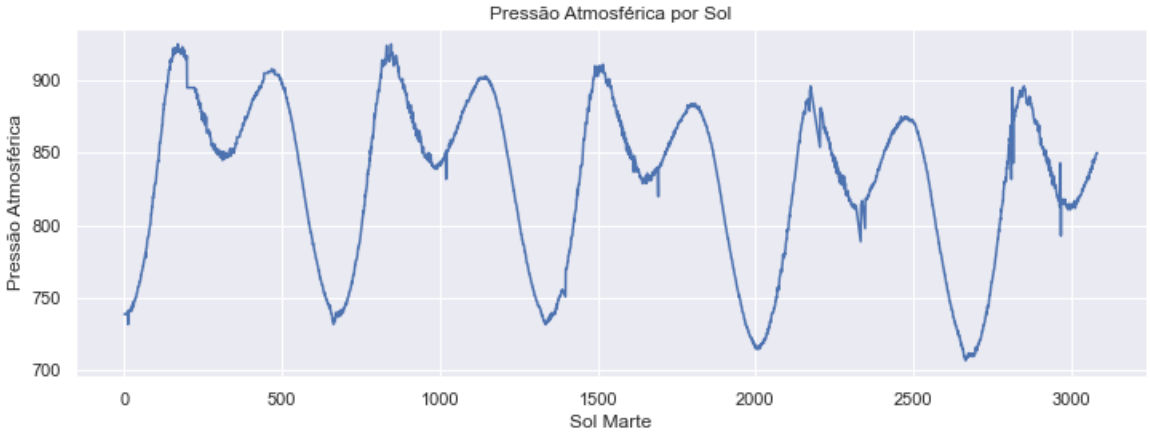

Figure 15: Pressão Atmosférica por Sol Marciano

Dessa maneira, na figura 15 percebe-se vales e picos em determinados períodos com determinada frequência em Marte, ou seja, o que configura-se como ciclos. Consequentemente, tal acontecimento pode identificar algum tipo de estação entre tais ciclos no planeta marciano, do qual pode-se interpretar que, a cada aproximadamente 680 sois marcianos, um ano em marciano, a pressão atmosférica volta a repetir valores de 680 sois atrás.

Outro fenômeno interessante de ser avaliado, e que pode ser visualizado na figura 15, é de que, sempre que um novo ciclo a cada ano marciano, os picos das pressões tenderam a ser inferiores aos picos dos ciclos anteriores. Da mesma maneira, os vales tenderam da mesma maneira, ou seja, a cada ano os vales alcançaram valores inferiores se comparado ao vale do ciclo anterior.

Verificando os histogramas das variáveis de temperatura máxima do ar e do solo, pode-se tornar esclarecedor as temperaturas máximas mais frequentes nos dados coletados pelo rover no planeta marciano, dado o período em questão. Dessa maneira, os histogramas se desenvolveram da seguinte maneira.

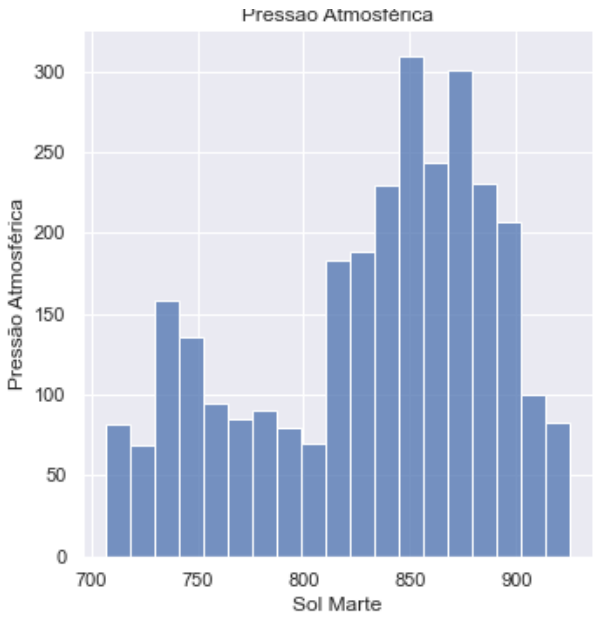

Figure 16: Histograma Pressão Atmosférica

Dessa forma, de acordo com a figura 16, percebe-se que a pressão atmosférica nos 2937 registros feitos pelo rover se concentraram no intervalo de 810 até 890 pascais, o que é condizente à ideia do curiosity estar explorando uma cratera, onde haverá mais pressão por ser um local mais profundo. Entretanto, pode-se inferir que a pressão mais frequente foi ade 850 pascais, da qual acumulou um frequência de mais de 300 sols marciano. Entretanto, percebe-se que a frequência de dias com pressão atmosférica entre 700 a 800 pascais foram as mais baixas.

Portanto, pode-se concluir que a pressão atmosférica majoritária em Marte, mais especificamente na cratera de Gale, se desenvolve no intervalo de 810 a 890 pascais. Entretanto, como explicado pelo Centro Astrobiologico (CAB), a pressão média geral de Marte varia em torno de 700 pascais. Dessa forma conclui-se o efeito da falta de atmosfera preenchida por gases no planeta vermelho, tornando-a até 160 vezes menor do que se comparado ao planeta Terra.

4.2.9 Nível UV

O índice ultravioleta (UV) mede o nível de radiação solar na superfície de determinado planeta. Quanto mais alto, maior o risco de danos à pele humana e de aparecimento de doenças como câncer. Portanto, o índice UV (índice ultravioleta) é um padrão internacional para a medição da intensidade de raios ultravioleta (UV) agindo sobre um determinado lugar levando-se em conta o tempo que incide.

Sendo assim, o nível de UV é um indicador da intensidade da radiação ultravioleta do Sol no local em que o Curiosity está explorando. A radiação ultravioleta é um agente prejudicial para a vida. Como explicado anteriormente, na Terra, a camada de ozônio impede que a luz ultravioleta prejudicial alcance a superfície, para o benefício de plantas e animais. No entanto, em Marte, devido à ausência de ozônio na atmosfera, a radiação ultravioleta atinge a superfície marciana.

Portanto, explorando os dados coletados da API do rover Curiosity, os dados referentes ao nível de radiação ultravioleta diário, em sol marciano, se desenvolveram da seguinte maneira no período aplicado:

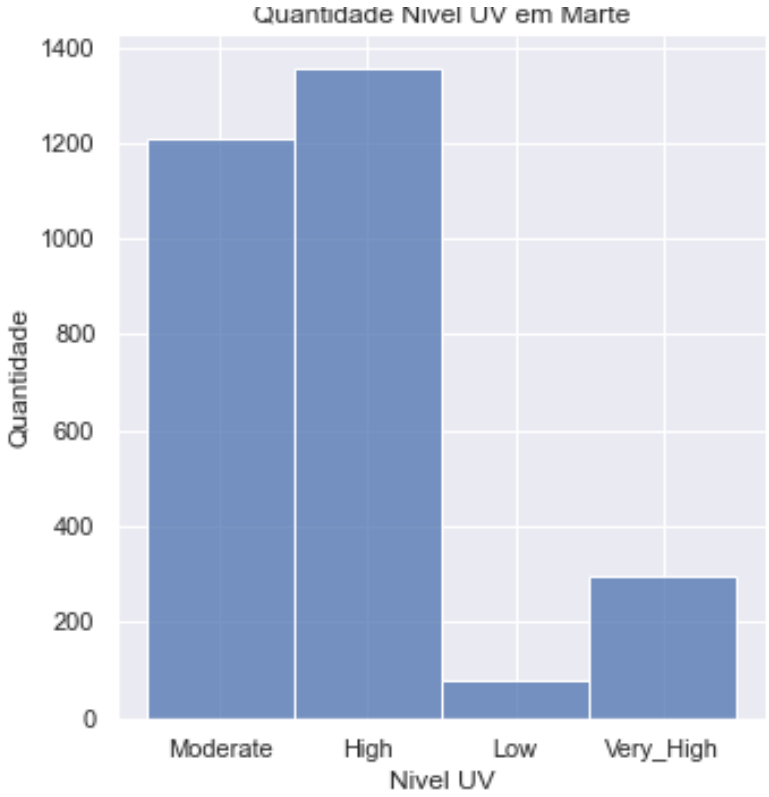

Figure 17: Nível UV por Sol Marciano

Os dados referentes ao nível da ação de raios ultravioletas no ambiente em que o Curiosity está inserido, foram catalogados pela CAB em quatro classes: muito alto, alto, moderado e baixo. Dessa maneira, visualizando a figura 17 percebe-se que, no período de coleta dos dados por parte do rover, houveram 1210 dias em que a radiação ultravioleta foi interpretada pelo rover como moderada, houveram 1354 dias em que a radiação ultravioleta foi interpretada como alta, 79 dias marcianos com radiação incidente sobre o rover interpretada como nível de radiação ultravioleta baixo, e, finalmente, 294 dias em que o rover Curiosity interpretou o nível de radiação ultravioleta como muito alto. Portanto, pode-se concluir que, nos 2937 dias de coleta dos dados, os níveis de radiação ultravioleta na cratera de Gale foram os níveis moderados e altos, situação condizente a de que marte não possui ozônio em sua atmosfera para proteger-se de tais raios.

A fim de entender a influência do nível de radiação ultravioleta nas temperaturas máximas, dedicou-se a criação de gráficos para tornar a análise mais visível ao leitor. À vista disso, gerou-se os seguintes gráficos:

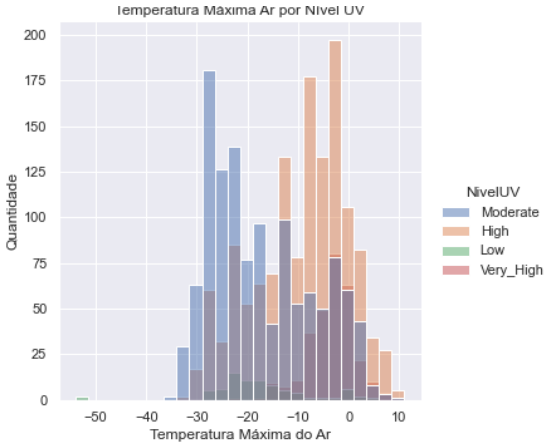

Figure 18: Temperatura Máxima do Ar por Nível UV

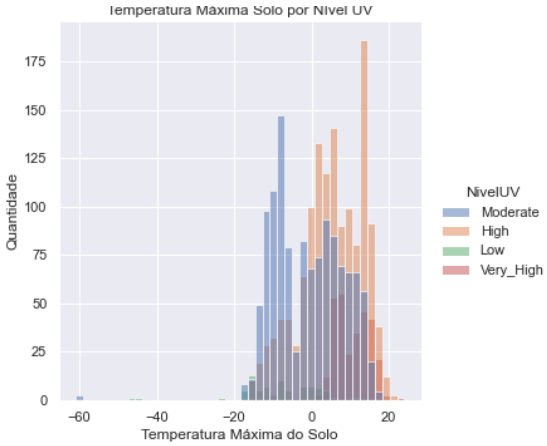

Figure 19: Temperatura Máxima do Solo por Nível UV

Portanto, de acordo com as figuras 18 e 19, percebe-se que, tanto para o ambiente do solo quanto do ar, quando a temperatura máxima foi maior que a mediana, dias que foram interpretados como radiação ultravioleta alta foram mais frequentes. Da mesma maneira, quando a temperatura máxima foi menor que a mediana, o nível de radiação ultravioleta, dias que foram interpretados como radiação ultravioleta moderada foram mais frequentes. Entretanto, isso quer dizer apenas que foi o mais frequente, sendo assim, é de se levar em consideração que, quando houve temperaturas máximas altas também foram captadas radiações de níveis moderada, baixa ou muito alto.

Consequentemente, pode ser levado em consideração que o nível de radiação ultravioleta possa influenciar as temperaturas máximas dos ambientes do solo e ar. Para verificar essa possível correlação entre as variáveis, posteriormente, será aplicada a matriz de correlação de Pearson, para verificar a veracidade de tal fenômeno em Marte. Entretanto, possivelmente a variável nível UV deverá ser convertida para tipo numérico, uma vez que a correlação de Pearson é calculado apenas para variáveis quantitativas, e não qualitativas, como é o caso atual de tal variável no presente momento.

4.2.10 Correlação

Após a etapa da análise exploratória de cada variável ser concluída, é valido explorar a relação de tais variáveis entre si no conjunto de dados proposto pela API do rover Curiosity. Dessa maneira, será avaliada a correlação das variáveis para entender a influência das mesmas.

A correlação, da qual será usada nesse projeto será a correlação de Pearson, mais especificamente a matriz de correlação de Pearson. De maneira simples, a correlação de Pearson mede a força do relacionamento linear entre duas variáveis quantitativas. Interpretando a tal matriz de correlação infere-se da seguinte maneira tais intervalos: um intervalo de -1 a 1, quando igual a -1 a correlação é linearmente negativa, quando igual a 1 a correlação é linearmente positiva e se igual a 0, a correlação é nula.

À vista disso, a aplicação do artificio estatístico de correlação em tal projeto se torna uma abordagem interessante, visto que existem diversas variáveis quantitativas na base de dados. Portanto, após aplicar tal abordagem, gerou-se a seguinte matriz de correlação de Pearson:

Figure 20: Matriz de Correlação de Pearson

Com a matriz de correlação representada na figura 20, entende-se que as temperaturas máximas do ar e do solo possuem correlação forte igual a 0,81, ou seja, em um gráfico de dispersão com dados dessas duas variáveis, haveria um tendência linear positiva. Entretanto, é interessante analisar que tais temperaturas máximas possuem correlação linearmente moderada positiva com as temperaturas mínimas do solo e do ar, fenômeno que também ocorre quanto avaliada a correlação das temperaturas mínimas com as temperaturas máximas.

Ao analisar a variável referente à pressão atmosférica, percebe-se que a mesma aparentemente possui correlação nula em relação às outras presentes no banco de dados. Sendo assim, infere-se que a pressão não interfere diretamente, seja positivamente ou negativamente, em relação às variáveis correspondentes às temperaturas, tanto as mínimas quanto às máximas do solo e do ar.

Dessa maneira, ao levar em consideração a segunda pergunta de negócio feita na etapa de entendimento de negócio, já é possível começar a criar caminhos até o modelo de regressão linear. Como solicitado a previsão da temperatura máxima do solo marciano, deve-se levar em consideração às correlações das variáveis temperatura máxima do ar, temperatura mínima do ar e temperatura mínima do solo, uma vez que essas, como avaliado na matriz, podem influenciar linearmente a temperatura máxima do solo.

4.3 Relatório Climático Marte

Após concluir a análise exploratória, dedicou-se a criação final desse relatório a fim de responder a primeira pergunta norteadora. Dessa maneira, esse relatório servirá como resumo e conclusão da análise exploratória feita, trazendo as informações gerais do possível clima meteorológico encontrado em Marte pelo rover Curiosity. Então, a seguir estão as informações que compõe o relatório climático de Marte de acordo com os dados da API com histórico coletado pelo rover Curiosity no período de 07 de Agosto de 2012 a 07 de Abril de 2021.

Como explicado em momentos anteriores, é válido ressaltar que o planeta Marte possui uma atmosfera bastante rasa, ou seja, uma camada tênue da qual contêm os gases ali presentes, assim afetando diretamente o clima e ambiente do planeta. Foi realizado, há alguns anos, estudos dos quais sugerem que, em algum momento do passado, Marte poderia ter tido características químicas e climáticas parecidas com a do planeta Terra. Dessa maneira, tentar entender o clima presente no planeta marciano faz-se justo para entender a história do sistema solar do qual os seres humanos se encontram.

Então, levando em consideração todos os dados coletados até então pelo rover Curiosity em sua jornada na cratera de Gale e a análise exploratória efetuada, entende-se que o ambiente na cratera de Marte seria majoritariamente frio. Como visto nesse projeto, as temperaturas do solo e do ar são diferentes devido à falta de atmosfera no planeta. Sendo assim, por conta disso, como citado pela CAB, a falta de atmosfera em relação à temperatura marciana pode se desenvolver como imaginar que está presente no Equador Marciano ao meio-dia, então você sentiria o verão em seus pés, mas o inverno em sua cabeça. Portanto, tais temperaturas, seja mínimas ou máximas, eram diferentes, entretanto, majoritariamente, pode-se considerar às temperaturas frias para um ser humano.

É interessante ressaltar que o planeta marciano possui solo arenoso e ambiente desértico, o que pode identificar grande amplitude térmica em relação ao solo, pois ambientes desérticos possuem grande capacidade de absorção do calor de dia, porém quando o sol se põe o clima torna-se frio. Da mesma forma, solos arenosos acabam por terem maior porosidade, havendo um menor contato entre as partículas do solo, dificultando assim o processo de condução do calor. Sendo assim, o ambiente solo em Marte possui grande amplitude térmica diante da falta de atmosfera e por possuir solo arenoso.

Em relação a pressão atmosférica, Marte possui uma pressão atmosférica em média 160 vezes menor do que em relação ao planeta Terra com uma média geral de 700 Pascais. Entretanto, como o rover Curiosity está localizado na cratera de Gale, a pressão atmosférica captada pelo dispositivo REMS é maior, em torno de 800 pascais, uma vez que crateras são mais profundas que o solo propriamente dito. Dessa maneira, a pressão atmosférica em Marte é menor do que em relação a do planeta Terra devido ao motivo da atmosfera marciana ser bastante tênue.

Ao que diz respeito ao histórico dos dados referentes ao nível de radiação ultravioleta, como explicado anteriormente, o rover Curiosity, por meio do dispositivo REMS interpreta o nível da radiação ultravioleta em 4 categorias. Como analisado, radiações de categorias moderada e alta foram as mais frequentes nos dias em que o rover coletou tais dados. Consequentemente, a falta de uma atmosfera com ozônio no planeta vermelho, faz com que a radiação ultravioleta consiga atingir a superfície marciana, diferente do planeta Terra do qual possui uma atmosfera com ozônio, repelindo a maioria dos respectivos raios ultravioletas.

Percebe-se então, que a atmosfera possui um papel muito importante em relação a regulação térmica de um planeta. Sendo assim, a atmosfera também tem o papel de conter o calor irradiado pelo planeta. Da mesma maneira, a atmosfera contém gases dos quais são denominados gases de efeito estufa, que garantem que parte do calor que chega ao planeta fique retido no planeta em questão.

Conclui-se que Marte possui a falta de uma atmosfera densa, ou seja, a existência de uma carência de gases, principalmente os de efeito estufa natural. Em conjunto com o solo arenoso, o clima de Marte é afetado diretamente em características como temperatura, pressão e o nível de radiação recebido na superfície marciana. Dessa forma, a absorção de calor torna-se pífia, tornando Marte um planeta majoritariamente frio e nocivo para o ser humano.

5 Preparação dos Dados

A etapa de preparação dos dados consiste em preparar os dados para a aplicação da modelagem no futuro. Sendo assim, nessa etapa serão aplicados, caso necessário, a seleção, limpeza, construção e integração dos dados. Entretanto, visando o escopo do projeto, será aplicada apenas a seleção e limpeza dos dados.

Como explicado anteriormente em 4.2.1, foi-se necessário preparar os dados para a análise exploratória, uma vez que todos esses estavam com o tipo diferente do qual o recomendado para análises estatísticas. Então, o tipo dos dados se desenvolveu da seguinte maneira.

- **DataTerra:** após a conversão do tipo da variável, o tipo tornou-se **datetime**.
- **DataMarte:** após a conversão do tipo da variável, o tipo tornou-se **inteiro**.
- **MaxTempAr:** após a conversão do tipo da variável, o tipo tornou-se **inteiro**.
- **MaxTempSolo:** após a conversão do tipo da variável, o tipo tornou-se **inteiro**.
- **MinTempAr:** após a conversão do tipo da variável, o tipo tornou-se **inteiro**.
- **MinTempSolo:** após a conversão do tipo da variável, o tipo tornou-se **inteiro**.
- **Pressao:** após a conversão do tipo da variável, o tipo tornou-se **inteiro**.
- **NivelUV:** até o momento, esta variável não necessitou a conversão do tipo.

Da mesma forma, em 4.2.9 foi citado a necessidade de transformar a variável Nivel UV em uma variável quantitativa, uma vez que, aparentemente em 4.2.9, existia correlação do nível de radiação ultravioleta com as temperaturas máximas em ambos ambientes, tanto o solo quanto ao ar. Portanto, resolveu-se aplicar tal recomendação, e então a variável Nível UV pode ser classificada numericamente da seguinte forma:

- **Low:** agora é representado pelo número **1**.
- **Moderate:** agora é representado pelo número **2**.
- **High:** agora é representado pelo número **3**.
- **Very_high:** agora é representado pelo número **4**.

Como explicado anteriormente na etapa de entendimento do negócio, os dados após tratados foram alocados em um banco de dados. O banco de dados utilizado para a reserva dos dados coletados via API e tratados via Python serão os bancos de dados PostegreSQL e SQLite. A escolha pelo banco de dados PostegreSQL se deu por ser referência no mercado de banco de dados, já em relação a utilização do SQLite se deu por tornar mais viável ao usuário chegar nos mesmos dados e análises feitas nesse projeto sem precisar das credenciais do autor no Postegre. Ambos bancos de dados são abertos ao público para a utilização.

Portanto, aplicando a transformação dessa variável do tipo qualitativo para quantitativo, será possível analisar novamente a correlação de Pearson, por meio da matriz de correlação, e então, de maneira mais clara, verificar a hipótese feita anteriormente em 4.2.9, de que existe correlação das temperaturas máximas com o nível de radiação ultravioleta captado pelo rover no planeta vermelho.

Então, após demonstrar as conversões concluídas dos tipos das variáveis aos respectivos tipos solicitados, pode-se concluir a etapa de preparação dos dados, configurando a finalização de uma das etapas mais impactantes em um projeto de mineração de dados do qual usa como metodologia o CRISP-DM. Caso essa etapa não seja concluída com o devido cuidado, o modelo de regressão pode não ser aplicável pelo motivo da tipagem das determinadas da variável estarem diferentes dos quais o modelo aceita.

5.1 Seleção dos Dados

Parte da etapa de seleção dos dados é identificar e selecionar os dados que podem alimentar o modelo de regressão linear da qual será construído posteriormente na etapa de modelagem. Dessa maneira, é necessário aplicar técnicas estatísticas para que faça sentido a seleção de tais dados, e então o modelo consiga atender todas as devidas exigências feitas anteriormente na etapa de entendimento do negócio e assim ser validado para a produção.

Como citado em 4.2.10, a técnica estatística da correlação de Pearson, e posteriormente da geração da matriz de correlação, se torna competente ao utilizar para a seleção dos dados. A utilização da correlação para avaliar quais dados serão selecionados e que possuem coesão, uma vez que, serão avaliadas as variáveis independentes que mais influenciam e impactam na variável-alvo.

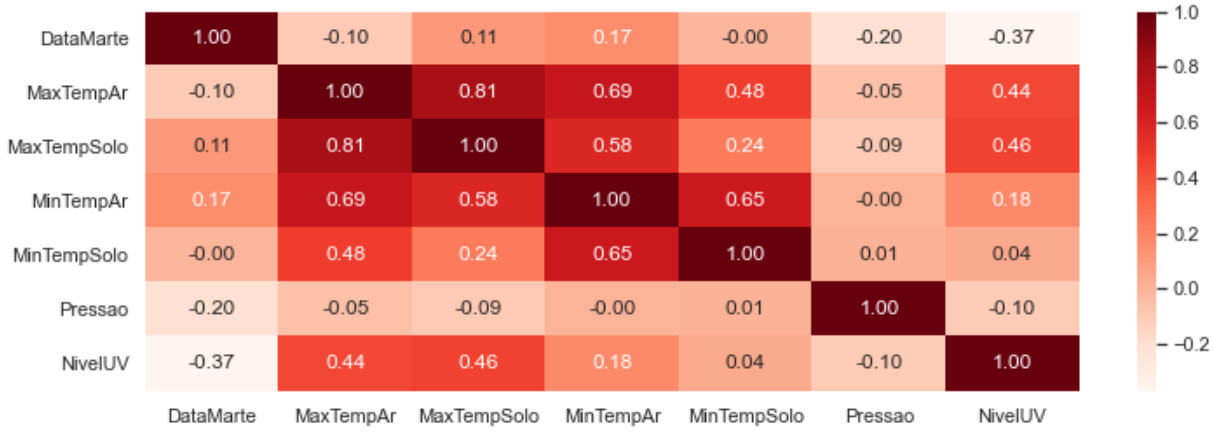

Figure 21: Matriz de Correlação de Pearson

Em modelos de regressão múltipla é necessário determinar um subconjunto de variáveis independentes que melhor explique a variável resposta, isto é, dentre todas as variáveis explicativas disponíveis, devemos encontrar um subconjunto de variáveis importantes para o modelo. Portanto, pode-se concluir que a variável Nivel UV afeta moderadamente temperaturas máximas, seja do solo ou do ar. Dessa maneira, como pode-se perceber na figura 21, após a transformação da variável Nível UV para uma variável quantitativa, as variáveis que mais impactam à temperatura máxima do solo são:

- **Temperatura Máxima do Ar**
- **Temperatura Mínima do Ar**
- **Nível UV**

Portanto, a seleção dos dados encontra-se concluída, uma vez que foram encontradas as variáveis explicativas das quais mais fazem influência em relação à variável alvo, ou seja a temperatura máxima do solo. Sendo assim, o modelo de regressão linear multivariado, do qual será criado posteriormente, será guiado pelas variáveis independentes citadas anteriormente.

5.2 Tratamento outliers

Outliers são dados que não são condizentes com o histórico da base de dados, ou seja, são basicamente "pontos fora da curva". Da mesma maneira, um outlier, por ser um ponto fora da curva, pode acabar causando anomalias, seja na estatística, como presenciado na figura 8, ou posteriormente, na etapa de geração e implementação de um modelo de machine learning. Portanto, o outlier é um dado mentiroso (dado o histórico dos dados) e deve ser tratado pois pode vir a causar análises e predições erradas caso utilizado-os.

Como percebeu-se nas figuras 9, 10, 12 e 12, existem outliers nos dados, dos quais não foram explicados pelo CAB porém, como anteriormente feito, foram levantadas algumas hipóteses para tentar explicar a inconsistência de alguns desses dados. Dessa forma, faz-se necessário o tratamento de tais dados que não condizem com o histórico para a posterior etapa de implementação do modelo de machine learning.

Sendo assim, para que os dados apresentem melhor consistência, aplicou-se a seguinte regra de negócio: tais dados deveriam ser filtrados em um intervalo de maneira que o dado mentiroso não tenha tanta relevância e posteriormente dê viés ao modelo. Dessa forma, após aplicar os filtros em determinado intervalo, para as variáveis que possuíam outliers, encontra-se o seguinte gráfico das mesmas:

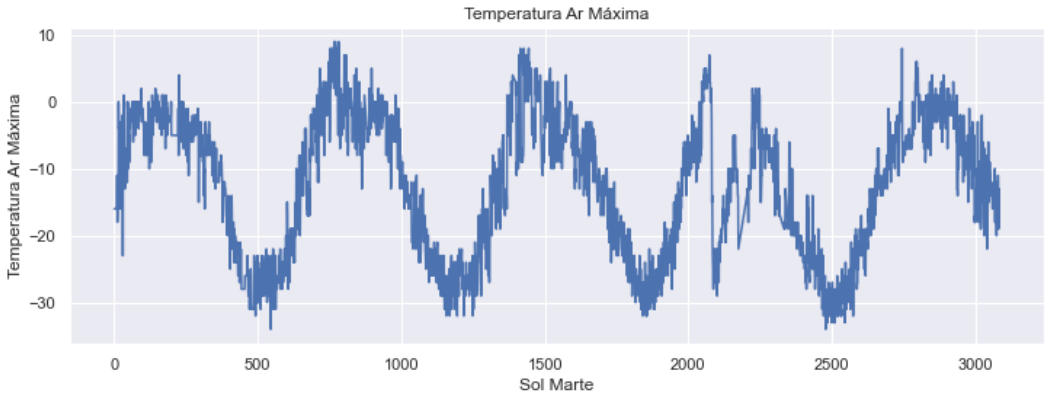

Figure 22: Temperatura Máxima Ar por Sol Marciano (Após remoção outlier)

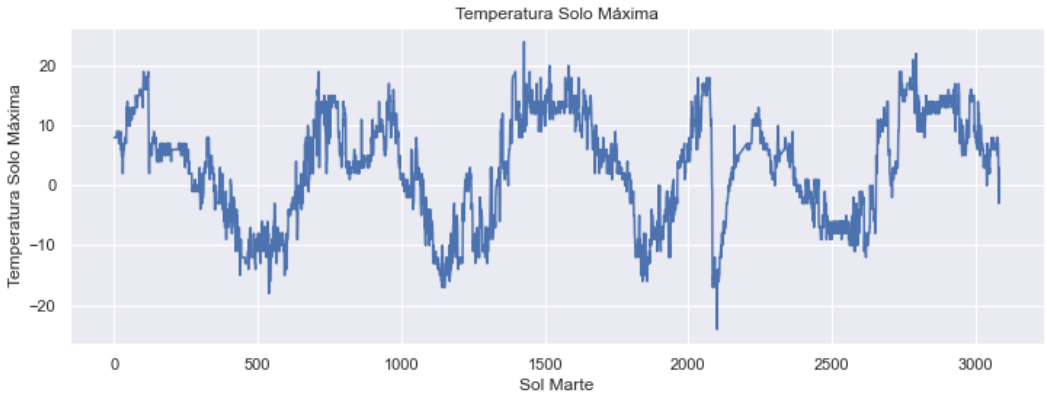

Figure 23: Temperatura Máxima Solo por Sol Marciano (Após remoção outlier)

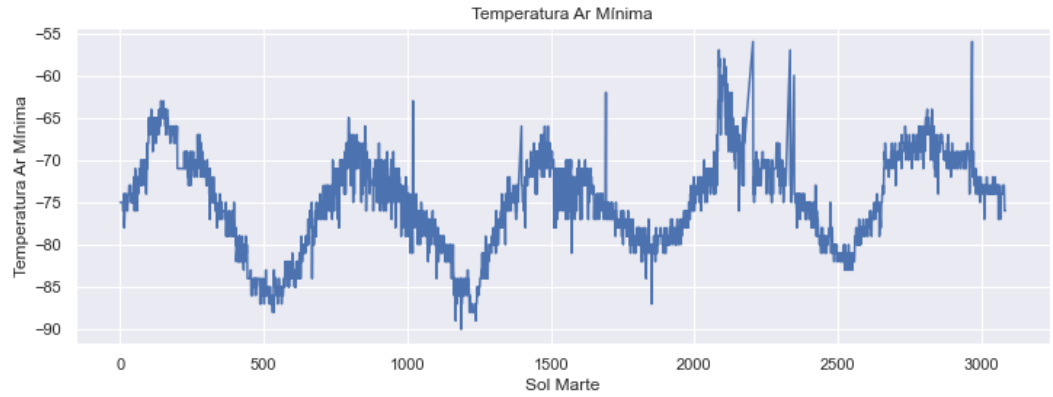

Figure 24: Temperatura Mínima Ar por Sol Marciano (Após remoção outlier)

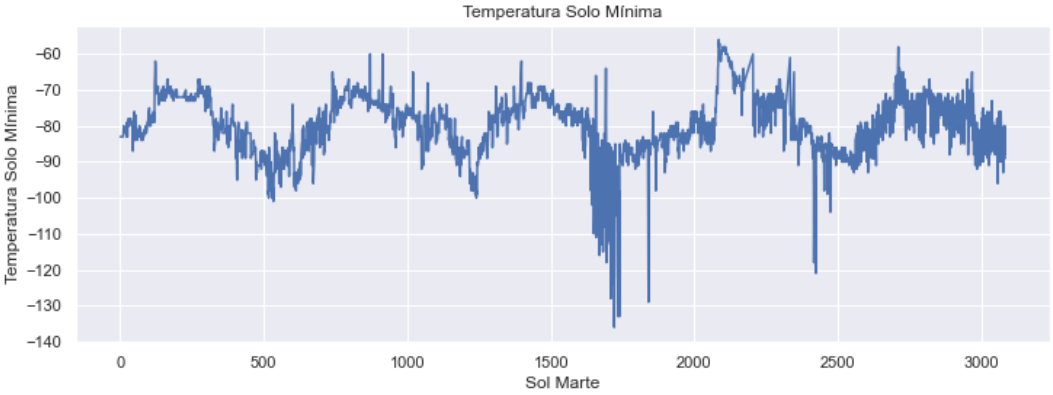

Figure 25: Temperatura Mínima Solo por Sol Marciano (Após remoção outlier)

Portanto, após analisar as figuras anteriores, e comparando-as com as figuras 9, 10, 12 e 12, percebe-se que boa parte dos outliers foram tratados de acordo com a regra de negócio, da qual filtrou-se os dados em um intervalo que não continha dados mentirosos.

Portanto, a etapa de tratamento dos dados, principalmente na manipulação de outliers, é considerada uma das mais valiosas ao que diz em relação a preparação do dados para a implementação do modelo de machine learning. Então, a etapa de implementação do modelo de machine learning pode ser desenvolvida com mais certeza, pois todos os cuidados, tratamentos e verificações foram analisados e aplicados na base de dados, podendo ter mais certeza em relação ao dado que será previsto, uma vez que o modelo estará sendo alimentado por dados limpos, condizentes e consistentes em relação ao histórico de tais.

6 Modelagem

A etapa de modelagem dos dados tem como entregue a escolha e a construção do algoritmo visando concluir a pergunta norteadora discutida inicialmente na etapa de entendimento do negócio.

Portanto, neste projeto será utilizado o modelo de regressão linear múltipla. A aplicação da regressão Múltipla será feita por meio da linguagem de programação Python utilizando principalmente a biblioteca scikit-learn.

6.1 Regressão Linear Múltipla

A regressão linear múltipla pode ser considerada como uma coleção de técnicas estatísticas para construir modelos que descrevem de maneira razoável relações entre várias variáveis explicativas de um determinado processo. Ou seja, de maneira simples, o modelo de regressão faz a atribuição de um valor contínuo a um elemento. Sendo assim, o modelo estatístico para a regressão múltipla pode ser representado, **generalizadamente**, pela seguinte fórmula:

$$Y = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon_i, \quad i = 1, ..., n$$

Após feita a introdução da fórmula matemática da qual será utilizada pelo modelo de previsão, torna-se justa a explicação das variáveis presente em tal, tornando assim o modelo mais claro. Sendo assim, tem-se:

- $x_{i1}, x_{i2}, ..., x_{ip}$**:** valores das variáveis explicativas, constantes conhecidas
- $\beta_0, \beta_1, \beta_2, ..., \beta_p$**:** os parâmetros ou coeficientes da regressão
- ϵ_i**:** erros aleatórios independentes

Da mesma forma, pode-se utilizar a abordagem matricial, ou seja, uma abordagem da fórmula de regressão anterior por meio de matrizes e vetores. Portanto, representa-se com a seguinte fórmula:

$$Y = X\beta + \epsilon$$

Sendo assim, a fórmula anterior pode ser denominada como Modelo Linear Geral. Vale ressaltar que nesse projeto não será efetuada análises de resíduos e diagnóstico de normalidade, pois tal projeto teve o intuito de introduzir o autor nos temas de regressão e suas aplicações em Python.

6.2 Modelo Scikit-learn

A biblioteca Scikit-learn é uma das principais bibliotecas na linguagem Python para machine learning. Nela, é possível encontrar ferramentas simples e extremamente eficientes para análises preditivas, como regressões por exemplo.

Dessa maneira, será necessário trabalhar com as variáveis explicativas que são as escolhidas para o modelo: Temperatura Máxima do Ar, Temperatura Mínima do Ar, Temperatura Mínima do Solo e Nível UV para prever um valor referente à Temperatura Máxima do Solo em Marte. Sendo assim, os dados necessitam ser separados em treino e teste, para que, posteriormente, seja avaliado e validado o modelo de regressão aplicado.

Portanto, a base de dados, será separada em 70% treino e 30% para efetuar testes de validação da regressão (erro médio quadrático, desvio médio quadrático e erro médio absoluto). Com o modelo definitivamente separado em treino e teste e finalizada a etapa de treinamento, é possível seguir para a próxima fase de avaliação e validação da regressão linear múltipla desenvolvida.

Importante ressaltar que, o código utilizado para a criação da regressão linear múltipla estará ao fim desse documento.

7 Avaliação

A fase de avaliação serve como a perícia do objetivo alcançado, avaliando se o modelo criado possui as características requisitadas na fase de entendimento do negócio. Sendo assim, como designado anteriormente, o modelo de previsão da expectativa de vida utilizando regressão linear múltipla deve possuir valores entre 0.3 e 5 para as métricas de MSE, RMSE e MAE.

7.1 Erro Médio Quadrático - MSE

O erro quadrático médio é definido como sendo a média da diferença entre o valor do estimador e do parâmetro ao quadrado. Desa maneira, o MSE, como é conhecido popularmente, é extremamente afetado caso haja valores discrepantes em relação à media, os famosos outliers.

7.2 Desvio Médio Quadrático - RMSE

O RMSE representa a raiz quadrada do segundo momento amostral, ou seja, o desvio padrão, das diferenças entre os valores previstos e os valores observados ou a média quadrática dessas diferenças. De forma simples, o RMSE é a raiz quadrada do MSE.

7.3 Erro Médio Absoluto - MAE

Em estatística, o erro médio absoluto, ou MAE como é conhecido popularmente, é uma medida de erros entre observações pareadas que expressam o mesmo fenômeno. Dessa maneira, o MAE não é afetado diretamente com valores discrepantes em relação à média.

7.4 Validação do Modelo

Após aplicar a regressão linear múltipla no conjunto de dados, levando em consideração as variáveis independentes e a variável alvo, como explicação em 6, encontrou-se as seguinte métricas referentes ao MSE, RMSE e MAE:

- **Erro Médio Quadrático (MSE):** aproximadamente 24 de score.
- **Raiz do Erro Médio Quadrático (RMSE):** aproximadamente 5 de score.
- **Erro Médio Absoluto (MAE):** aproximadamente 4 de score.

Dessa maneira, como é possível declarar, apenas as métricas RMSE e MAE atingiram o objetivo proposto alcançando o score entre 0.3 a 5. Então, basta entender o motivo do qual o MSE foi tão superior se comparado às métricas MAE e RSME. Como explicado anteriormente, o Erro Médio Quadrático sofre bastante quando há valores discrepantes em relação à media, e, como informado na fase de análise exploratória, tais variáveis explicativas possuíam valores discrepantes em determinados momentos em relação à media. Dessa forma, tal acontecimento pode ter afetado diretamente no score MSE.

Portanto, após avaliar a utilização da métrica de MSE, constatou-se que não seria a melhor métrica a ser utilizada para validação da regressão linear múltipla criado, uma vez que tal métrica é afetada pelos valores discrepantes. Porém, ao utilizarmos o RMSE, do qual é o Desvio Médio Quadrático, então entende-se que valores discrepantes não serão camuflados, uma vez que será retornado a raiz quadrada dos dados, pois o desvio padrão calcula o "erro" se quiséssemos substituir um dos valores coletados pelo valor da média.

Dessa forma, foi levado em consideração apenas as métricas RMSE e MAE, pois são as mais recomendadas para a validação da regressão linear múltipla criada. Então, as métricas RMSE e MAE estão no intervalo solicitado anteriormente, ou seja, estão entre 0.3 e 5. Ou seja, ao prever a temperatura máxima do solo, pode ocorrer no máximo até 5 graus Celsius de erro. Sendo assim, o modelo de regressão criado pode ser considerado como validado, uma vez também que testes foram efetuados, comprovando a proximidade dos scores com os resultados requeridos.

8 Implementação

A fase de implementação é a última no ciclo da metodologia de projeto de mineração de dados do CRISP-DM. Essa fase tem como objetivo iniciar e finalizar a produção do modelo e sua implementação.

A implementação do modelo criado nesse projeto será efetuada por meio da biblioteca Streamlit programada em Python. Com o uso do Streamlit será possível criar um ambiente interativo com o usuário, para que o mesmo forneça os valores para as variáveis preditoras e o modelo retorne um valor contínuo referente à variável alvo, no caso a temperatura máxima do solo em Marte.

8.1 Streamlit

Streamlit é uma biblioteca Python de código aberto que facilita a criação de aplicações web personalizadas para diversos campos, como por exemplo em aprendizado de máquina e ciência de dados. Sendo assim, esse projeto a utilizará para fazer a implementação das conclusões das perguntas norteadoras feitas anteriormente na etapa de entendimento de negócio.

Então, a implementação web será aplicada em três etapas, ou seja, haverá três páginas das quais poderão ser acessadas pelo menu do qual será criado. Na primeira página haverá um breve resumo do projeto e seus escopos, bem como os links contendo o código utilizado e essa documentação da qual está sendo efetuada. Já a segunda página conterá o relatório climático de Marte, finalizado na etapa de análise exploratória, do qual conseguirá responder a primeira pergunta formulada na etapa de entendimento do negócio. Por fim, a terceira página irá conter o modelo de regressão linear múltipla, mostrando um breve resumo ao usuário de como funciona tal aplicação e também opções de interação do usuário com o modelo criado, fazendo assim o usuário passar dados que alimentarão o modelo e devolverá o resultado previsto para a temperatura máxima do solo marciano de acordo com os dados disponibilizados pelo usuário na interação. A implementação do modelo pode ser acessada **clicando aqui**.

9 Conclusão

Como explicado anteriormente, o código utilizado para a criação desse projeto estará disponível no fim dessa documentação.

Vale ressaltar que, como explicado anteriormente, a inserção e o acesso aos dados direto do banco de dados requer dados sigilosos dos autores. Sendo assim, utilizou-se o SQLite para a implementação em Streamlit, pois quando o projeto estiver no repositório do GitHub, o usuário consiga acessar os mesmos dados utilizados sem a necessidade de acessar o banco de dados dos autores. Portanto, os arquivos utilizados para hospedar os resultados no Streamlit estão utilizando a consulta em um banco de dados SQLite para preservar a senha do banco de dados PostegreSQL.

Portanto, a última etapa da metodologia CRISP-DM está concluída, dispondo então de todas as outras etapas anteriores validadas e finalizadas. Sendo assim, todas as perguntas norteadoras foram respondidas e implementadas em um modelo em produção dando a possibilidade do resultado estar mais próximo ao usuário final. Então, pode-se concluir que as indagações e projeções solicitadas conseguiram atender todos os resultados esperados, sendo validadas por métricas, quando necessário.

9.1 Próximos passos

Como acontece em todo projeto, sempre há algo a ser melhorado. Da mesma forma, os próximos passos desse projeto englobaria continuar a coleta dos dados oriundos do instrumento REMS presente no rover Curiosity e a melhora no modelo de regressão linear multivariada para a previsão da temperatura máxima do solo, da qual posteriormente poderá conter mais variáveis para alimentar o modelo.

10 Referências

Centro de Astrobiología (CAB), Rover Environmental Monitoring Station Mars Science Laboratory. Disponível em: http://cab.inta-csic.es/rems/en#previous-sol.

Hedibert Freitas, Análise de Regressão Linear Múltipla I. Disponível em: http://hedibert.org/wp-content/uploads/2014/02/Econometria201401-Aula04-ARLM-I-Estimacao.pdf.

Javier Gomez-Elvira, National Aeronautics and Space Administration. Disponível em: https://mars.nasa.gov/msl/spacecraft/instruments/rems/.

Natasha Romanzoti, O que a Curiosity encontrou no solo de Marte. Disponível em: https://hypescience.com/confira-os-resultados-da-primeira-analise-completa-de-curiosity-sobre-o-solo-marciano/.

National Aeronautics and Space Administration, MARS Curiosity Rover Weather. Disponível em: https://mars.nasa.gov/msl/weather/.

National Aeronautics and Space Administration, MARS Curiosity Rover. Disponível em: https://mars.nasa.gov/msl/home/.

National Aeronautics and Space Administration, MARS Exploration Program. Disponível em: https://mars.nasa.gov/.

National Aeronautics and Space Administration, Mars Science Laboratory/Curiosity. Disponível em: https://mars.nasa.gov/msl/news/pdfs/MSL_Fact_Sheet.pdf.

National Aeronautics and Space Administration, There's Water on the Moon?. Disponível em: moon.nasa.gov/news/155/theres-water-on-the-moon.

Novakovic, B., Senenmut: An Ancient Egyptian Astronomer. Disponível em: https://ui.adsabs.harvard.edu/abs/2008POBeo..85...19N/abstract.

Portal Action (EstatCamp), REGRESSÃO LINEAR MÚLTIPLA. Disponível em: http://www.portalaction.com.br/analise-de-regressao/regressao-linear-multipla.

Paulo Cesar Sentelhas e Luiz Roberto Angelocci, Temperatura do ar e do solo. Disponível em: http://www.leb.esalq.usp.br/leb/aulas/lce306/Aula6_2012.pdf.

R. Orosei, S. E. Lauro, E. Pettinelli, E. Pettinelli, A. Cicchetti, M. Coradini and B. Cosciotti, Radar evidence of subglacial liquid water on Mars. Disponível em: https://science.sciencemag.org/content/361/6401/490.

11 Código utilizado

Nesta parte da documentação estará o código utilizado para a criação do projeto. Entretanto, informações sensíveis como por exemplo, o acesso ao banco de dados dos autores não estarão disponíveis.

```
 # Data from Mars Curiosity - 2012 ~ 2021

import json                    #Capturar dados Json
import requests                #Fazer requerimento na API
import pandas as pd            #Tratamento e Modelagem dos dados
import seaborn as sns          #Criação de Gráficos
import sqlalchemy                #Insercao e acesso ao BD
import matplotlib.pyplot as plt #Criação de Gráficos


###############ACESSANDO A API###################################
url = "https://mars.nasa.gov/rss/api/?feed=weather&category=msl&feedtype=json"
data = requests.get(url).json()
#print(json.dumps(data, indent=4))

######################Criando as variaveis#################
####################
terraData = []

for earthDate in data['soles']:
    terraData.append(earthDate['terrestrial_date'])

####################
maiorTempAr = []

for marsMaxTempAir in data['soles']:
    maiorTempAr.append(marsMaxTempAir['max_temp'])

####################
maiorTempSolo = []

for marsMaxTempGround in data['soles']:
    maiorTempSolo.append(marsMaxTempGround['max_gts_temp'])

####################
menorTempAr = []

for marsMinTempAir in data['soles']:
    menorTempAr.append(marsMinTempAir['min_temp'])

####################
menorTempSolo = []

for marsMinTempGround in data['soles']:
    menorTempSolo.append(marsMinTempGround['min_gts_temp'])

####################
```

```
pressao = []

for marsPress in data['soles']:
    pressao.append(marsPress['pressure'])

####################
marteData = []

for marsDate in data['soles']:
    marteData.append(marsDate['sol'])

####################
radiacaoMarte = []

for marsUV in data['soles']:
    radiacaoMarte.append(marsUV['local_uv_irradiance_index'])

####################
#Criando um dicionario para as variaveis criadas,
#entao posteriormente sera feita um data frame usando pandas
dicionario_geral = {'DataTerra': terraData, 'DataMarte': marteData, 'MaxTempAr': maiorTempAr,
                    'MaxTempSolo': maiorTempSolo,'MinTempAr': menorTempAr,
                    'MinTempSolo': menorTempSolo,'Pressao': pressao, 'NivelUV': radiacaoMarte}
df = pd.DataFrame.from_dict(dicionario_geral)
df.head()

## Transformando o tipo das variáveis
#Convertendo DataTerra para tipo Data
df['DataTerra'] = pd.to_datetime(df['DataTerra']).dt.tz_localize(None)

#Convertendo DataMarte para tipo Inteiro
df[['DataMarte']] = df[['DataMarte']].astype(int).astype('Int64')

#Convertendo MaxTempSolo para tipo Inteiro
df[['MaxTempSolo']] = df[['MaxTempSolo']].replace('--', None)
df[['MaxTempSolo']] = df[['MaxTempSolo']].astype(int).astype('Int64')

#Convertendo MinTempAr para tipo Inteiro
df[['MinTempSolo']] = df[['MinTempSolo']].replace('--', None)
df[['MinTempSolo']] = df[['MinTempSolo']].astype(int).astype('Int64')

#Convertendo MaxTempAr para tipo Inteiro
df[['MaxTempAr']] = df[['MaxTempAr']].replace('--', None)
df[['MaxTempAr']] = df[['MaxTempAr']].astype(int).astype('Int64')

#Convertendo MinTempAr para tipo Inteiro
df[['MinTempAr']] = df[['MinTempAr']].replace('--', None)
df[['MinTempAr']] = df[['MinTempAr']].astype(int).astype('Int64')

#Convertendo Pressao para tipo Inteiro
df[['Pressao']]= df[['Pressao']].replace('--', None)
df[['Pressao']] = df[['Pressao']].astype(int).astype('Int64')

#Convertendo NivelUV '--' para None
```

```
df[['NivelUV']] = df[['NivelUV']].replace('--', None)

#Monitorando
df.dtypes
```

Nesse momento estariam as linhas de código utilizadas para enviar o data frame para o banco de dados e posteriormente acessá-lo. Entretanto, como explicado anteriormente, são informações sensíveis e não será cedido.

```
#Análise Exploratória
df['DataTerra'] = pd.to_datetime(df['DataTerra']).dt.tz_localize(None)
df.info()
df.describe()

###################Gráficos###################

####Correlacao####
sns.set(rc={'figure.figsize':(12,4)})
sns.heatmap(df.corr(),
            annot = True,
            fmt = '.2f',
            cmap='Reds')
plt.savefig("correlacaoAntes.png")

#Ajustando Seaborn
sns.set(rc={'figure.figsize':(12,4)})
sns.set_theme(style="darkgrid")

##Testes gráficos
sns.set(rc={'figure.figsize':(10, 8)})
sns.pairplot(df)

sns.set(rc={'figure.figsize':(10, 8)})
sns.pairplot(df, hue='NivelUV')

#########Graficos Temperaturas Máximas#########
###Ar
sns.lineplot(data = df, x="DataMarte", y="MaxTempAr").set(title='Temperatura Ar Máxima',
                                                          xlabel='Sol Marte',
                                                          ylabel='Temperatura Ar Máxima')
plt.savefig("maxTempAr.png")

####Solo####
sns.lineplot(data = df, x="DataMarte", y="MaxTempSolo").set(title='Temperatura Solo Máxima',
                                                            xlabel='Sol Marte',
                                                            ylabel='Temperatura Solo Máxima')
plt.savefig("maxTempSolo.png")

####Ar####
sns.displot(df, x="MaxTempAr").set(title='Histograma Temperatura Ar Máxima',
                                   xlabel='Temperatura Ar Máxima', ylabel='Frequência')
plt.savefig("maxTempArHist.png")

####Solo####
```

```
sns.displot(df, x="MaxTempSolo").set(title='Histograma Temperatura Solo Máxima',
                                     xlabel='Temperatura Solo Máxima', ylabel='Frequência')
plt.savefig("maxTempSoloHist.png")


#########Graficos Temperaturas Mínimas#########
####Ar####
sns.lineplot(data = df, x="DataMarte", y="MinTempAr").set(title='Temperatura Ar Mínima',
                                                          xlabel='Sol Marte',
                                                          ylabel='Temperatura Ar Mínima')
plt.savefig("minTempAr.png")

####Solo####
sns.lineplot(data = df, x="DataMarte", y="MinTempSolo").set(title='Temperatura Solo Mínima',
                                                            xlabel='Sol Marte',
                                                            ylabel='Temperatura Solo Mínima')
plt.savefig("minTempSolo.png")

####Ar####
sns.displot(df, x="MinTempAr").set(title='Histograma Temperatura Ar Mínima',
                                   xlabel='Temperatura Ar Mínima', ylabel='Frequência')
plt.savefig("minTempArHist.png")

####Solo####
sns.displot(df, x="MinTempSolo").set(title='Histograma Temperatura Solo Mínima',
                                     xlabel='Temperatura Solo Mínima', ylabel='Frequência')
plt.savefig("minTempSoloHist.png")

#########Graficos Pressao#########
sns.lineplot(data = df, x="DataMarte", y="Pressao").set(title='Pressão Atmosférica por Sol',
                                                        xlabel='Sol Marte',
                                                        ylabel='Pressão Atmosférica')
plt.savefig("pressao.png")

sns.displot(data = df, x="Pressao").set(title='Pressão Atmosférica por Sol',
                                        xlabel='Sol Marte', ylabel='Pressão Atmosférica')
plt.savefig("pressaoHist.png")

#########Graficos NivelUV#########
sns.displot(data = df, x="NivelUV").set(title='Quantidade Nivel UV em Marte',
                                        xlabel='Nivel UV', ylabel='Quantidade')
plt.savefig("NivelUV.png")

sns.displot(data = df, x="MaxTempAr",
            hue = 'NivelUV').set(title='Temperatura Máxima Ar por Nível UV',
                                 xlabel='Temperatura Máxima do Ar',
                                 ylabel='Quantidade')
plt.savefig("maxTempArNivelUV.png")

sns.displot(data = df, x="MaxTempSolo",
            hue = 'NivelUV').set(title='Temperatura Máxima Solo por Nível UV',
                                 xlabel='Temperatura Máxima do Solo',
                                 ylabel='Quantidade')
plt.savefig("maxTempSoloNivelUV.png")
```

```
#########Manipulacao #########
df.NivelUV.value_counts()
mapping_dictionary = {"NivelUV":{ "Low": 1, "Moderate": 2, "High": 3, "Very_High": 4}}
df = df.replace(mapping_dictionary)
df.head()


#########Regressao Linear Multipla#########
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error
from sklearn.metrics import mean_absolute_error

import numpy as np

X = df[['MaxTempAr', 'MinTempAr', 'MinTempSolo', 'NivelUV']]
Y = df['MaxTempSolo']

model = LinearRegression() #criando a variavel pra usar reg linear

#separando os dados para treino e teste
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.3, random_state=1) =
model.fit(X_train, Y_train)#treinando o modelo

y_test_predicted = model.predict(X_test)
#y_test_predicted.shape

print("MSE: {}".format(mean_squared_error(Y_test, y_test_predicted)))

print("RSME: {}".format(mean_squared_error(Y_test, y_test_predicted, squared = False)))

print("MSA: {}".format(mean_absolute_error(Y_test, y_test_predicted)))

print("="*50)

new_array = np.array([-13, -76, -89, 2]).reshape(-1, 4)
print('MARTE: ', model.predict(new_array))
```