**SARS-CoV-2 Variant Evolution Analysis** project:

**Project Objectives**

1. **Data Collection and Metadata Analysis:**
   * Retrieve SARS-CoV-2 genome sequences and associated metadata (e.g., host, collection date, geographic location) from public repositories like NCBI Virus or GISAID.
   * Curate and organize the metadata to identify trends and key attributes relevant to variant analysis.
2. **Mutation Identification:**
   * Perform multiple sequence alignment to detect mutations across the genomes, focusing on key regions like the Spike protein.
   * Identify and catalog the most prevalent and unique mutations, including their potential impact on viral function.
3. **Variant Classification:**
   * Use lineage classification tools (e.g., Pangolin) to assign genomes to known SARS-CoV-2 variants.
   * Investigate the distribution of variants across different geographic regions and timeframes.
4. **Phylogenetic Analysis:**
   * Construct phylogenetic trees to infer evolutionary relationships among SARS-CoV-2 genomes.
   * Use phylogenetic insights to track the emergence and spread of variants globally.
5. **Functional Annotation:**
   * Annotate mutations using tools like SnpEff or Ensembl VEP to predict their potential impact on protein structure and function.
   * Highlight mutations of interest, such as those linked to increased transmissibility, immune escape, or drug resistance.
6. **Data Visualization:**
   * Create informative visualizations, including:
     + Mutation frequency plots.
     + Geographic distribution maps.
     + Phylogenetic trees showing evolutionary trajectories.
   * Utilize tools like Nextstrain for interactive visual representation.
7. **Reporting and Insights:**
   * Summarize findings in a comprehensive report or presentation.
   * Provide actionable insights on how the observed mutations and variants might impact public health, vaccine efficacy, and therapeutic strategies.

**Week 1: Data Collection and Metadata Organization**

* **Goal:** Gather and organize all required data for downstream analysis.
* **Tasks:**
  + Register and download genome sequences and metadata from NCBI Virus or GISAID.
  + Filter data by location, collection date, and completeness.
  + Curate metadata into a structured format (CSV or TSV) for analysis.
* **Tools:** NCBI Virus portal, NCBI Datasets CLI, Python (for data parsing).
* **Deliverables:** A curated dataset of genome sequences and metadata.

**Week 2: Sequence Alignment and Mutation Analysis**

* **Goal:** Detect mutations and identify key genomic variations.
* **Tasks:**
  + Align genome sequences using MAFFT.
  + Use bcftools to call mutations.
  + Summarize common mutations and their locations in the genome.
* **Tools:** MAFFT, bcftools, Python.
* **Deliverables:**
  + Aligned genome sequences.
  + Mutation call file (VCF) and summary table.

**Week 3: Variant Classification**

* **Goal:** Assign genomes to known SARS-CoV-2 lineages and identify emerging variants.
* **Tasks:**
  + Use Pangolin to classify genomes into variants.
  + Analyze lineage distribution across metadata attributes (e.g., geography, time).
* **Tools:** Pangolin, Python (for metadata analysis).
* **Deliverables:**
  + Lineage classification results.
  + Lineage distribution plots.

**Week 4: Phylogenetic Analysis**

* **Goal:** Infer evolutionary relationships and spread of variants.
* **Tasks:**
  + Build phylogenetic trees using IQ-TREE or Nextstrain.
  + Interpret tree topology to identify variant origins and spread patterns.
* **Tools:** IQ-TREE, Nextstrain CLI.
* **Deliverables:**
  + Phylogenetic tree(s).
  + Geographic mapping of variant spread.

**Week 5: Functional Annotation**

* **Goal:** Predict the impact of mutations on viral proteins.
* **Tasks:**
  + Annotate mutations using SnpEff or Ensembl VEP.
  + Focus on mutations in critical regions like the Spike protein.
  + Identify mutations linked to immune escape or transmissibility.
* **Tools:** SnpEff, Ensembl VEP.
* **Deliverables:**
  + Annotated mutation table.
  + Highlighted list of impactful mutations.

**Week 6: Visualization and Reporting**

* **Goal:** Create a comprehensive report with visualizations and actionable insights.
* **Tasks:**
  + Generate visualizations (mutation frequency plots, phylogenetic trees, maps).
  + Write a detailed report summarizing methods, results, and implications.
  + Prepare a presentation or dashboard for sharing results.
* **Tools:** Python (Matplotlib, Seaborn), Nextstrain, Word/LaTeX.
* **Deliverables:**
  + Final report with visualizations.
  + Presentation or dashboard.

**Tips for Time Management:**

* **Set daily goals** based on the weekly tasks.
* Use **version control** (e.g., Git) to manage your files and scripts.
* Allocate time for troubleshooting technical issues.
* If short on time, focus on objectives that provide the most insights (e.g., phylogenetics and variant classification).

**Skills You'll Practice:**

* Remote data access and handling.
* Genome alignment and mutation analysis.
* Phylogenetic tree construction and evolutionary modeling.
* Data visualization using Python or R.
* Biological interpretation of genomic trends.

**Deliverables:**

A comprehensive report or interactive dashboard highlighting:

* Mutation patterns in SARS-CoV-2 variants.
* Evolutionary trajectories and spread across regions.
* Implications for vaccine and therapeutic developments.