Geothermal Energy Supply Curve Analysis (EGS 4k Limited Access, Moderate)

Bryant Le

2025-10-17

Introduction

Geothermal energy is a promising renewable energy source that can provide reliable electricity throughout the USA. This project analyzes a dataset from the Open Energy Data Initiative of various geothermal systems throughout the country.

The objective is to examine geothermal potential by location and infrastructure access, determine relationships between capacity, distance to transmission, and developable area, and identify which states have the highest potential for efficient geothermal development.

The dataset follows the DKIW framework: - Data: Raw EGS Site Information - Information: Cleaned and structured variables and visualizations. - Knowledge: Derived supply curves and state comparisons. - Wisdom: Actionable insights on geothermal prioritization.

Loading Data and Overview

```
geo_data <- read.csv("geo_egs_4k_limited_access_moderate_supply_curve.csv")
```

```
str(geo_data)
```

```
## 'data.frame':    18326 obs. of  10 variables:
##  $ sc_point_gid         : int  799 1179 1180 1559 1560 1566 1567 1949 2319 2320 ...
##  $ latitude             : num  49 48.9 48.9 48.8 48.8 ...
##  $ longitude            : num  -122 -122 -122 -122 -122 ...
##  $ country              : chr  "United States" "United States" "United States" "United States" ...
##  $ state                : chr  "Washington" "Washington" "Washington" "Washington" ...
##  $ county               : chr  "Whatcom" "Whatcom" "Whatcom" "Whatcom" ...
##  $ timezone             : int  -8 -8 -8 -8 -8 -8 -8 -8 -8 -8 ...
##  $ area_developable_sq_km: num  7.2625 3.7417 2.8991 0.0973 0.3958 ...
##  $ capacity_ac_mw       : num  17.94 8.99 7.56 0.23 1.06 ...
##  $ dist_spur_km         : num  1.44 4.78 11.19 15.58 21.38 ...
```

```
summary(geo_data)
```

```
##   sc_point_gid      latitude       longitude         country         
##  Min.   :  799   Min.   :25.88   Min.   :-123.74   Length:18326      
##  1st Qu.:21734   1st Qu.:33.68   1st Qu.:-115.38   Class :character  
##  Median :43630   Median :39.12   Median :-108.55   Mode  :character  
##  Mean   :44395   Mean   :39.07   Mean   :-107.70                     
##  3rd Qu.:66191   3rd Qu.:44.39   3rd Qu.:-101.99                     
##  Max.   :98222   Max.   :49.00   Max.   : -71.31                     
##     state              county             timezone     area_developable_sq_km
##  Length:18326       Length:18326       Min.   :-8.00   Min.   :3.240e-04     
##  Class :character   Class :character   1st Qu.:-8.00   1st Qu.:4.047e+00     
##  Mode  :character   Mode  :character   Median :-7.00   Median :1.905e+01     
##                                        Mean   :-6.97   Mean   :3.048e+01     
##                                        3rd Qu.:-6.00   3rd Qu.:4.976e+01     
##                                        Max.   :-5.00   Max.   :1.309e+02     
##  capacity_ac_mw       dist_spur_km    
##  Min.   :  0.00076   Min.   :  0.000  
##  1st Qu.: 12.73750   1st Qu.:  3.644  
##  Median : 59.86561   Median :  9.652  
##  Mean   : 99.76742   Mean   : 17.118  
##  3rd Qu.:158.07013   3rd Qu.: 22.225  
##  Max.   :685.33076   Max.   :332.319  
```

This data set includes the following variables:

- sc_point_gid: A unique identifier for each supply curve point (like a site ID).
- latitude, longitude: Spatial coordinates of the potential site. These coordinates allow mapping, spatial joins, and distance calculations.
- state, country, county: Jurisdiction information. Useful for aggregating supply potential by geography.
- timezone: Local time zone offset.
- area_developable_sq_km: The land area (square kilometers) that is actually usable for development after exclusions.
- capacity_ac_mw: Potential alternating current generation capacity (MW) that could be built on the developable land.
- dist_spur_km: Distance (km) to the nearest spur line (transmission connection point).

Each record represents a potential geothermal supply site.

Data Cleaning

It is important to check for missing values in the data set to prevent skewed results.

```r
geo_data %>%
  summarise(
    missing_latitude = sum(is.na(latitude) | latitude == ""),
    missing_longitude = sum(is.na(longitude) | longitude == ""),
    missing_state = sum(is.na(state) | state == ""),
    missing_county = sum(is.na(county) | county == ""),
    missing_country = sum(is.na(country) | country == ""),
    missing_timezone = sum(is.na(timezone) | timezone == ""),
    missing_area = sum(is.na(area_developable_sq_km) |
                                    area_developable_sq_km == ""),
    missing_capacity = sum(is.na(capacity_ac_mw) |
                               capacity_ac_mw == ""),
    missing_dist = sum(is.na(dist_spur_km) |
                             dist_spur_km == "")
  )
```

```
##   missing_latitude missing_longitude missing_state missing_county
## 1                0                 0          1958           1958
##   missing_country missing_timezone missing_area missing_capacity missing_dist
## 1            1958                0            0                0            0
```

Some of the values in the state, county, and country columns are missing. Renaming the missing values as "Unknown" preserves the original records for the upcoming visualizations.

```
geo_data <- geo_data %>%
  mutate(
    state  = ifelse(is.na(state) | state == "", "Unknown", state),
    county = ifelse(is.na(county) | county == "", "Unknown", county),
    country = ifelse(is.na(country) | country == "", "Unknown", country)
  )
```

```
geo_data %>%
  filter(state == "Unknown") %>%
  summarise(total_capacity = sum(capacity_ac_mw, na.rm = TRUE))
```

```
##   total_capacity
## 1       148415.9
```

Feature Engineering

Calculating the capacity density is beneficial to help identify which places offer the most geothermal energy per unit of land. This will help determine the potential and promote confidence before investing in these areas.

```
geo_data <- geo_data %>%
  mutate(capacity_density = capacity_ac_mw / area_developable_sq_km)
```

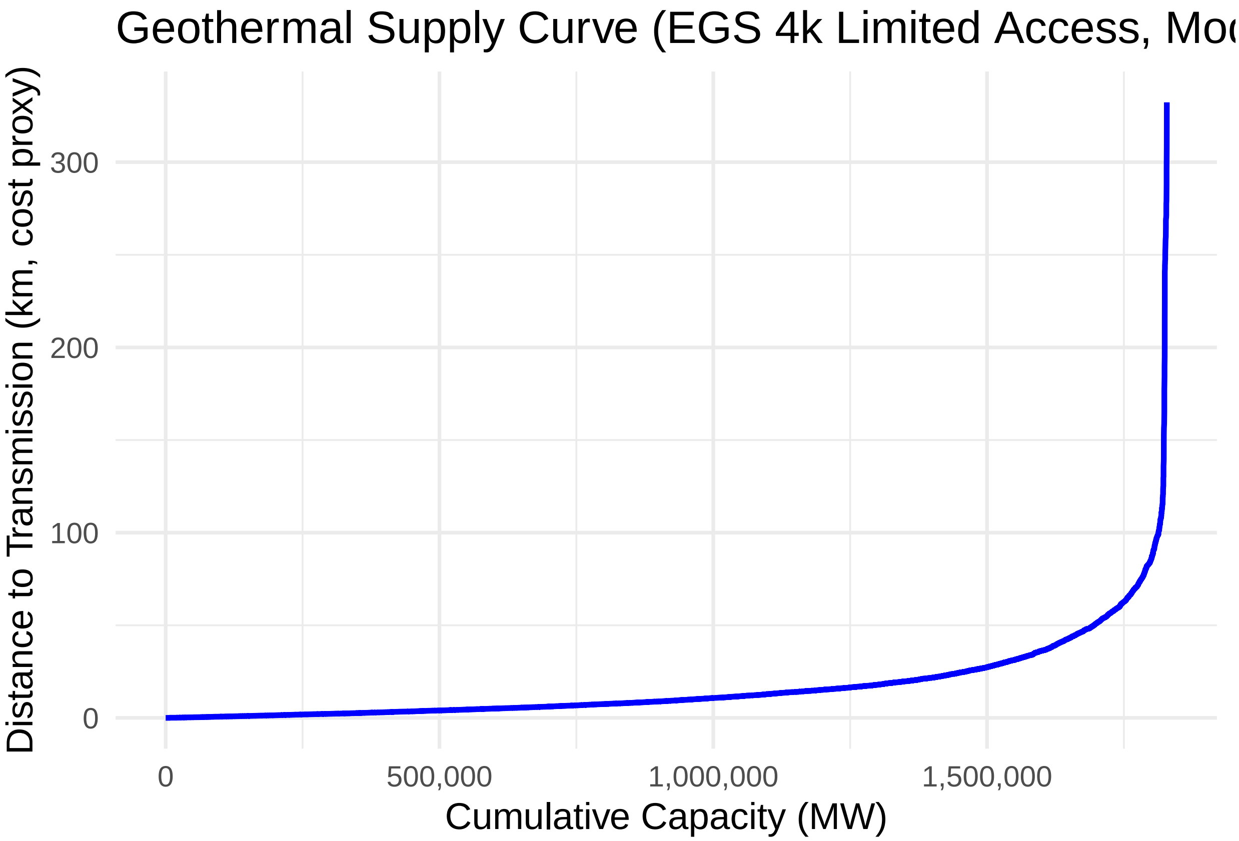
Exploratory Analysis of Supply Curve

Calculating the capacity_density will be important to compare the efficiency of geothermal utilization across various sites. Higher density means that more energy will be produced per unit of land.

It is imperative to examine how cumulative capacity will increase with distance to transmission. Ideally, a flatter curve would prove the area to be of high potential with low cost and risk in investment. However, a curved scale would suggest that the cheapest sites closest to transmission would become scarce, and expanding capacity would result in a significant increase in cost. Unfortunately, it appears that the latter is true and would make expansion difficult.

```
supply_curve <- geo_data %>%
  arrange(dist_spur_km) %>%
  mutate(cumulative_capacity = cumsum(capacity_ac_mw))
```

```
ggplot(supply_curve, aes(x = cumulative_capacity, y = dist_spur_km)) +
  geom_line(color = "blue", linewidth = 1) +
  labs(title = "Geothermal Supply Curve (EGS 4k Limited Access, Moderate)", x = "Cumulative Capacity (MW
  scale_x_continuous(labels = comma) +
  theme_minimal(base_size = 14)
```



Spatial Visualization of Supply Sites

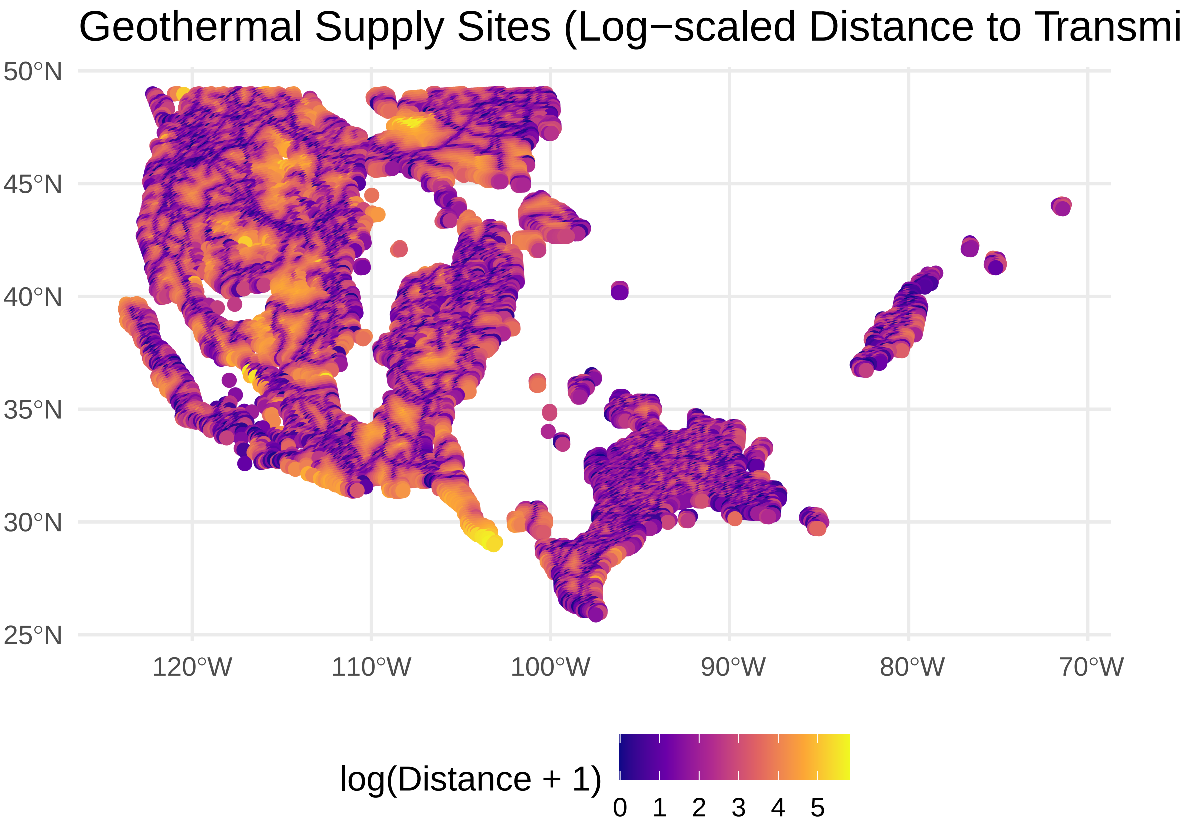
Given that the dataset contains GPS coordinates for each location, it will be beneficial to visualize the coordinates on a map. Since the data mostly appears to come from the USA, readability should not be a concerning issue.

The visualization suggests that the west of USA has the highest geothermal potential, but proximity appears to be inconsistent. Texas appears to have higher occurrences of short transmission proximity distances than the west of USA.

```r
geo_sf <- st_as_sf(
  geo_data,
  coords = c("longitude", "latitude"),
  crs = 4326
)
```

```r
ggplot() +
  geom_sf(
    data = geo_sf,
    aes(size = capacity_ac_mw, color = log1p(dist_spur_km)),
    alpha = 1,
    size = 2
  ) +
  geom_point() +
  scale_color_viridis_c(option = "plasma") +
  labs(title = "Geothermal Supply Sites (Log-scaled Distance to Transmission)",
       color = "log(Distance + 1)",
       size = "Capacity (MW)") +
  theme_minimal(base_size = 13) +
  theme(legend.position = "bottom")
```

```
## Ignoring unknown labels:
## * size : "Capacity (MW)"
```



State Data Comparison

It is important to compare the top state's geothermal capacity to determine which states have the most EGS potential. However, there is a large presence of "Unknown" locations that may skew the results. Therefore, it is necessary to create two charts: one with "Unknown" and one without "Unknown". For the supply curve analysis, "Unknown" will be omitted.

```
top_states <- geo_data %>%
  filter(state != "Unknown") %>%
  group_by(state) %>%
  summarise(total_capacity = sum(capacity_ac_mw, na.rm = TRUE)) %>%
  arrange(desc(total_capacity)) %>%
  slice(1:5) %>%
  pull(state)
```

```
supply_curve_state <- geo_data %>%
  group_by(state) %>%
  arrange(dist_spur_km, .by_group = TRUE) %>%
  mutate(cumulative_capacity = cumsum(capacity_ac_mw)) %>%
  ungroup()
```

```
supply_curve_state_top <- supply_curve_state %>%
  filter(state %in% top_states)
```

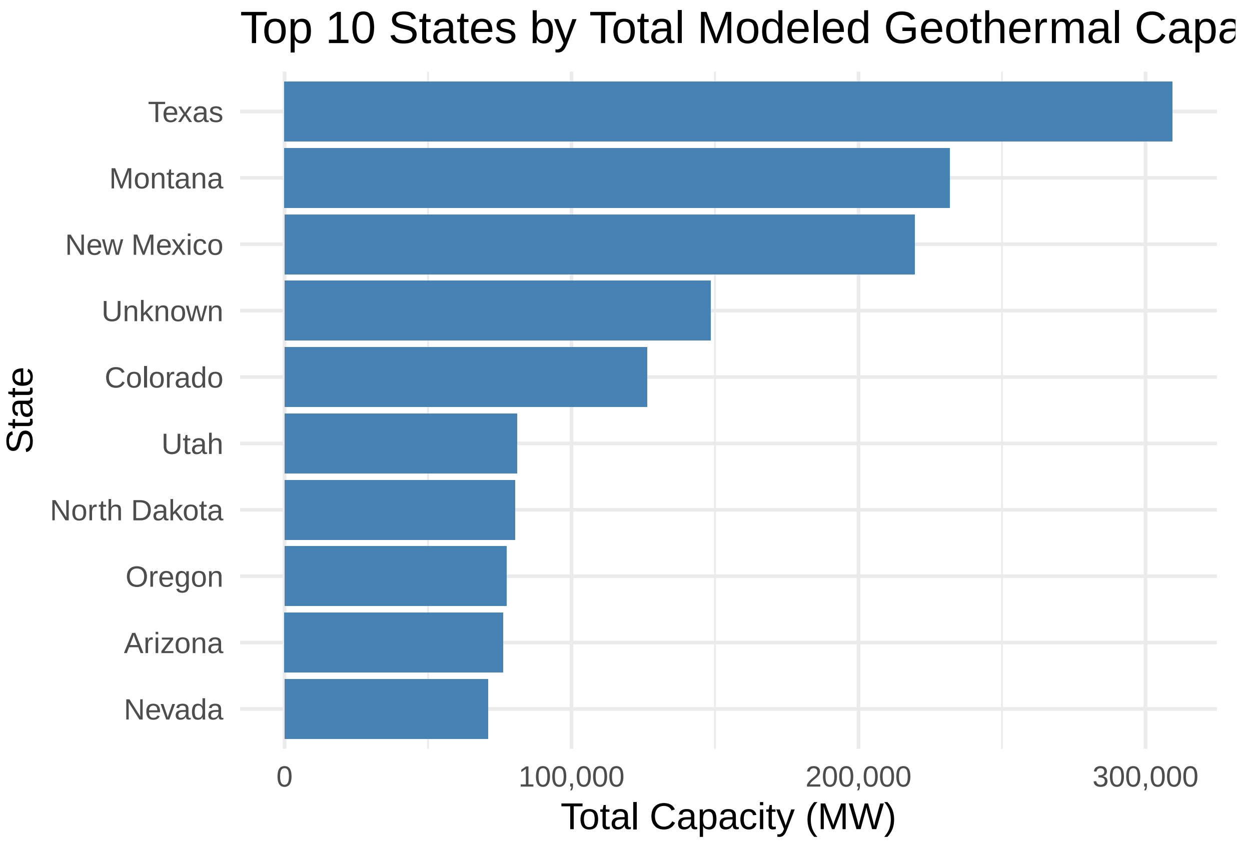
```
state_capacity_summary <- geo_data %>%
  group_by(state) %>%
  summarise(
    total_capacity_mw = sum(capacity_ac_mw, na.rm = TRUE),
    n_sites = n()
  ) %>%
  arrange(desc(total_capacity_mw))
```

Geothermal Capacity of Top States

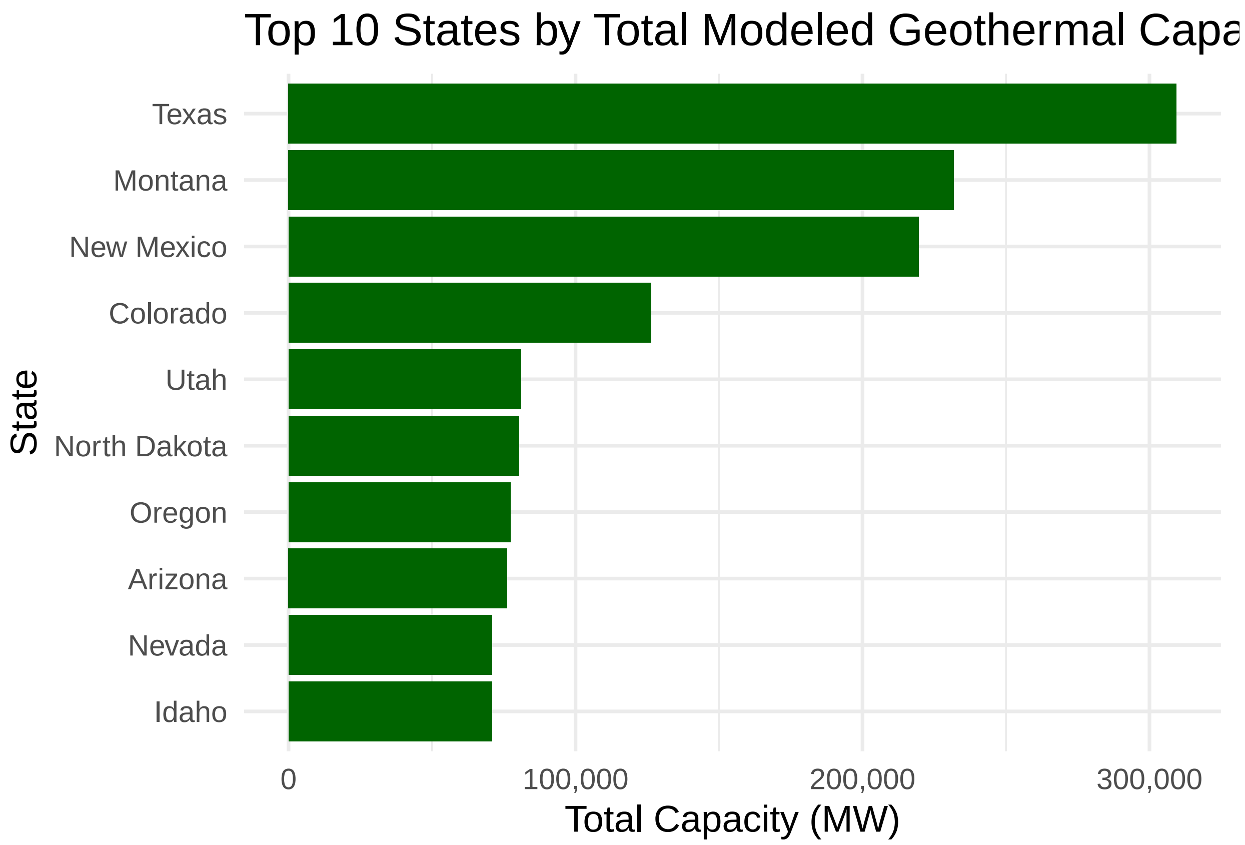
The following bar charts compare the total geothermal capacity across the states. The results display the following top 5 states: Texas, Montana, New Mexico, Unknown, and Colorado (Utah is fifth if "Unknown" is excluded). Analyzing the transmission distances will be the next step to determine which states would have the most EGS potential.

Previous results indicate that California has many geothermal points, but the EGS potential is noticeably lower based on the data set. This suggests that potential capacity is likely based on subsurface temperature instead of hydrothermal generation.

```
ggplot(state_capacity_summary %>%
         slice_max(total_capacity_mw, n = 10) %>%
         mutate(state = fct_reorder(state, total_capacity_mw)),
       aes(x = state, y = total_capacity_mw)) +
  geom_col(fill = "steelblue") +
  coord_flip() +
  labs(
    title = "Top 10 States by Total Modeled Geothermal Capacity (EGS 4k Limited Access)",
    x = "State",
    y = "Total Capacity (MW)"
  ) +
  scale_y_continuous(labels = scales::comma) +
  theme_minimal(base_size = 14)
```



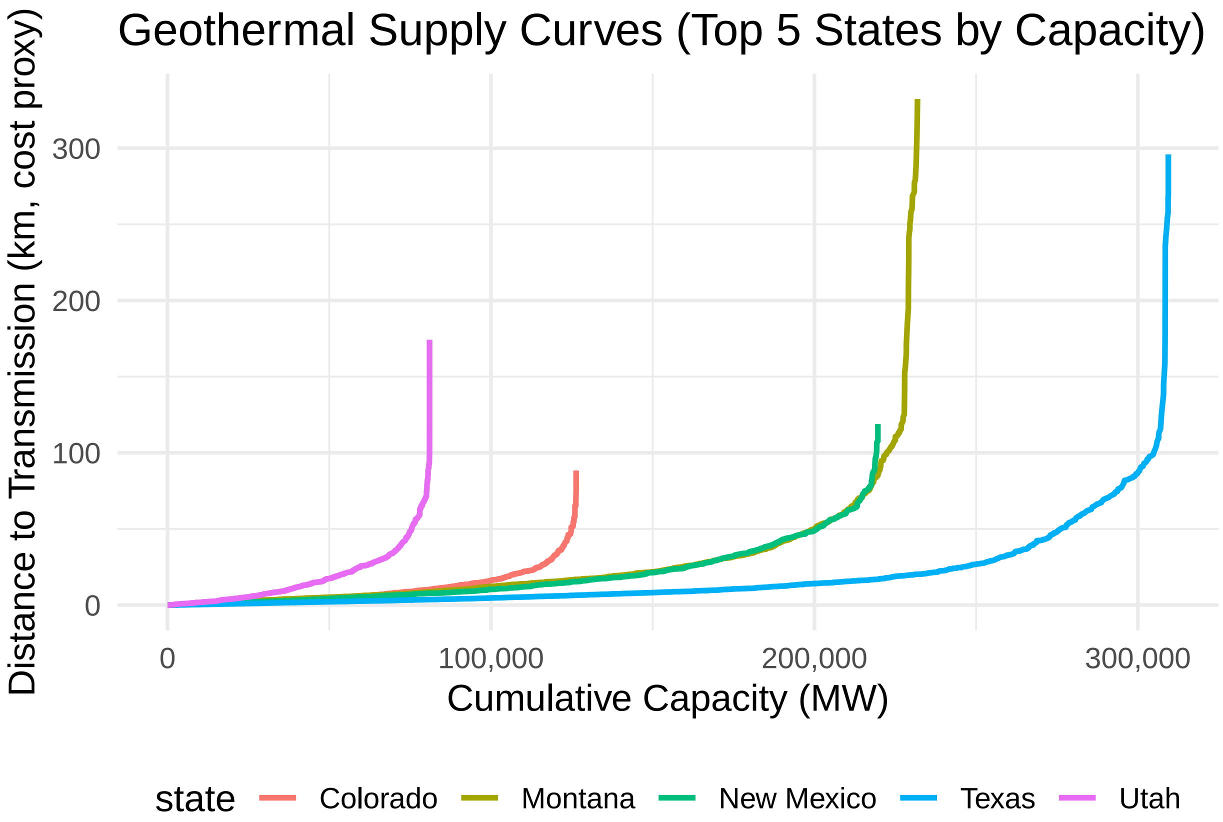
```
ggplot(state_capacity_summary %>%
         filter(state != "Unknown") %>%
         slice_max(total_capacity_mw, n = 10) %>%
         mutate(state = fct_reorder(state, total_capacity_mw)),
       aes(x = state, y = total_capacity_mw)) +
  geom_col(fill = "darkgreen") +
  coord_flip() +
  labs(
    title = "Top 10 States by Total Modeled Geothermal Capacity (Excluding Unknown)",
    x = "State",
    y = "Total Capacity (MW)"
  ) +
  scale_y_continuous(labels = scales::comma) +
  theme_minimal(base_size = 14)
```



Supply Curves for Top States

Supply curve analysis suggests there is large EGS potential in Texas and New Mexico due to their large cumulative capacities and low transmission distances. This suggests favorable infrastructure access and lower connection costs. While Montana and Utah have strong geothermal gradients, they would not be as optimal compared to Texas and New Mexico due to higher transmission distances and connection costs. Based on the results, Texas would be the strongest candidate for scalable geothermal deployment.

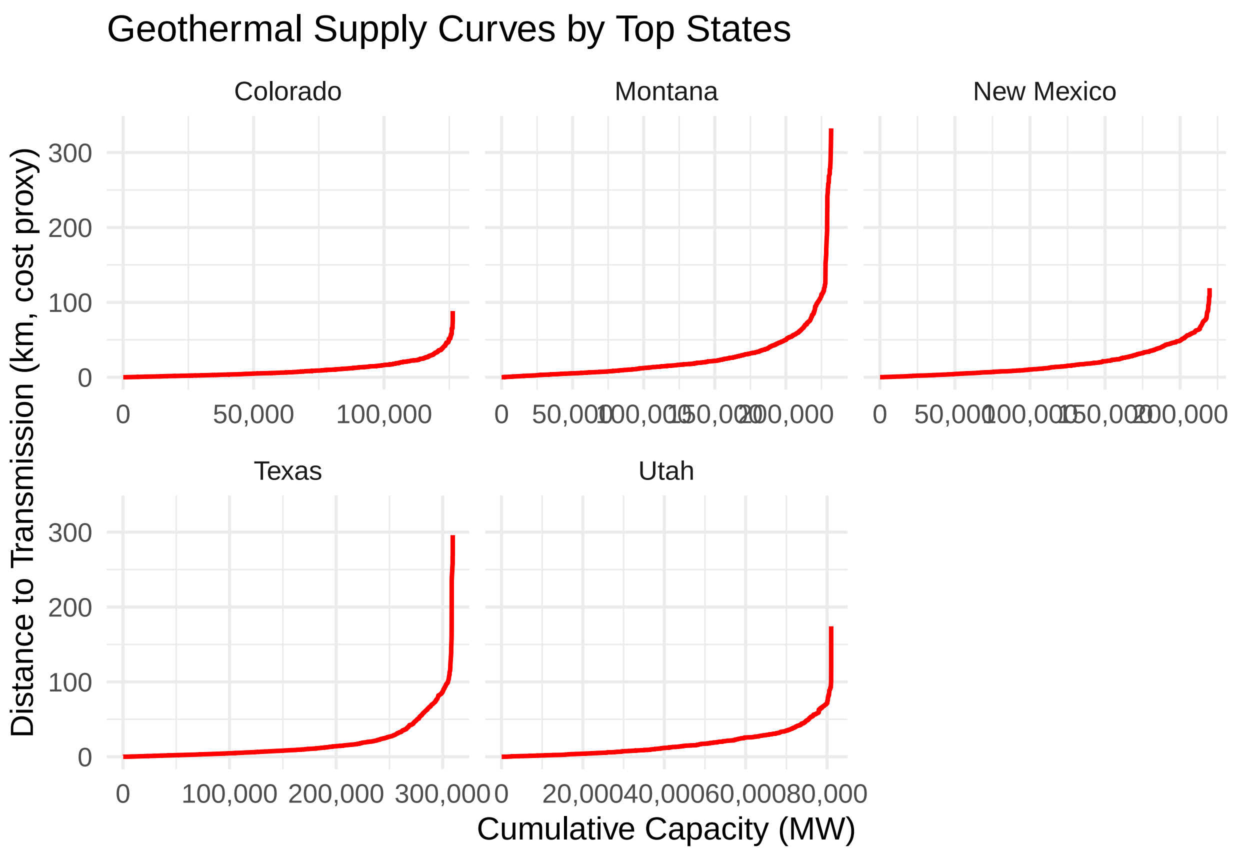
```r
ggplot(supply_curve_state_top,
       aes(x = cumulative_capacity, y = dist_spur_km, color = state)) +
  geom_line(linewidth = 1) +
  labs(title = "Geothermal Supply Curves (Top 5 States by Capacity)", x = "Cumulative Capacity (MW)", y
  scale_x_continuous(labels = comma) +
  theme_minimal(base_size = 14) +
  theme(legend.position = "bottom")
```



Faceted Comparison

This visualization enhances readability by providing a direct comparison of the curve shapes and ranges of the top 5 states by capacity.

```
ggplot(supply_curve_state_top,
       aes(x = cumulative_capacity, y = dist_spur_km)) +
  geom_line(color = "red", linewidth = 0.8) +
  facet_wrap( ~ state, scales = "free_x") +
  labs(title = "Geothermal Supply Curves by Top States", x = "Cumulative Capacity (MW)", y = "Distance t
  scale_x_continuous(labels = comma) +
  theme_minimal(base_size = 12)
```



Conclusion

The project cleaned and structured geothermal records to visualize cumulative and state-level supply curves to determine optimal regions for strategic EGS development investment.

After cleaning and engineering the EGS supply curve modeling data, the following insights were gained:

- Western states contain most of the potential sites, with Texas and New Mexico demonstrating the most potential due to lower transmission distances.
- Montana and Utah's geothermal gradients are significant, but they may not be optimal due to high infrastructure costs as transmission distance increases.
- "Unknown" contains 1958 missing state records, which will warrant geospatial cross-validation for further study and to determine the cause of the missing values.

Implementing cost modeling to determine the potential expense to pursue future projects may prove beneficial for future analysis.