

Intel Data Center



INTRODUCTION: Intel, the semiconductor manufacturing powerhouse, is planning on building a new data center. Energy availability and usage are some of the key considerations in deciding on a location of the data center. For example, which regions produce a surplus of energy, and are therefore more likely to provide energy at cheaper prices? Which regions rely more on renewable energy sources?

In this project, co-designed with Intel's Sustainability Team, you'll write SQL queries that will power your analysis and create visualizations that will help the Intel team select the best location for the new data center.

HOW IT WORKS: Follow the prompts in the questions below to investigate your data. Post your answers in the provided boxes: the **yellow boxes** for the queries you write, **purple boxes** for visualizations and **blue boxes** for text-based answers.

– Data Set **Descriptions**

In this project you'll query 3 datasets as well as write a query to generate a new dataset that you will use in your tableau visualizations. The `intel.energy_data` dataset will be the main dataset you'll be working with. The `intel.energy_by_plant` and `intel.power_plants` datasets will be joined for an in-depth analysis of energy production at the power plant level.

Read below to learn more about the datasets and their features.

intel.energy_data: Contains information about daily energy production and consumption for different regions in the United States.

- **balancing_authority** - A Balancing Authority is responsible for maintaining the electricity balance within its region. This is a company that makes sure electricity is being exchanged between electric providers and regions so that no region runs out of electricity due to high demand.
- **date** - The date the energy was produced.
- **region** - The electric service area within a geographic area of the USA. e.g. California, Midwest, etc.
- **time_at_end_of_hour** - The time and date after energy was generated, .e.g., energy generated between 1pm-2pm will show up as 2pm in this field.
- **demand** - The energy demand in megawatts (MW) on the grid (what the houses/business are using).
- **net_generation** - The energy produced in MW in the region by all sources e.g., wind, coal, nuclear, etc.
- **all_petroleum_products** - The energy produced in MW by petroleum products.
- **coal** - The energy produced in MW by all coal products
- **hydropower_and_pumped_storage** - The energy produced in MW by water power and pumped heat sources.
- **natural_gas** - The energy produced in MW by natural gas sources
- **nuclear** - The energy produced in MW from nuclear fuel sources
- **solar** - The energy produced in MW by solar panels and other solar energy capturing methods.
- **wind** - The energy produced in MW from wind turbines and other wind sources.

intel.power_plants: Contains general information about power plants in the United States.

- **plant_name** - The name of the power plant.
- **plant_code** - The unique identifier of the plant.
- **region** - The region in the US where the power plant is located. Matches the regions in the intel.energy_data
- **state** - The state where the power plant is located.
- **primary_technology** - The primary technology used to generate electricity at the power plant.

intel.energy_by_plant: Contains total energy production information at the plant for the year 2022.

- `plant_name` - The name of the power plant.
 - `plant_code` - The unique identifier of the plant.
 - `energy_type` - The kind of energy generated by the power plant. Either renewable energy or fossil fuel.
 - `energy_generated_mw` - The total energy generated, in MegaWatts, at the plant for the year 2022.
-

– Task 1: Energy Generation

Let's first identify regions that are net energy producers. Not all regions generate enough energy to meet the local demand. Some regions purchase power from other regions, while others sell their surplus to regions in need.

- A.** Write a query using the `intel.energy_data` table that calculates the sum total of energy produced, grouped by each region. Sort the output by highest total energy. Which region has the highest positive total energy?

HINT: Total energy is equal to the difference between `net_generation` and `demand`.

```
SELECT
    region,
    SUM(net_generation - demand) AS total_energy
FROM
    intel.energy_data
GROUP BY
    region
ORDER BY
    total_energy DESC;
```

The region that has the highest positive total energy is the Mid-Atlantic. It has 31693087 of total energy.

- B.** Intel is interested in regions that generate a large amount of energy from renewable sources. Renewable energy is defined as any energy generated from hydropower_and_pumped_storage, wind, and solar sources.

Write a query that calculates the sum total of renewable energy by region. Sort the output by the region with the highest renewable energy. What are the top two regions for total renewable energy production?

HINT: You need to add the 3 energy sources together in one line before doing your group by: `SUM(col1 + col2 + col3) AS new_column`

```
SELECT
    region,
    SUM(hydropower_and_pumped_storage + wind + solar) AS
total_renewable_energy
FROM
    intel.energy_data
GROUP BY
    region
ORDER BY
    total_renewable_energy DESC;
```

The top two regions for total renewable energy production are:

1. Northwest : 199266574
2. Texas : 131367234

- C. Modify your query slightly so that it calculates the **percentage** of renewable energy by region.

HINT: Divide the amount of renewable energy by the sum total of `net_generation`, and then multiply the result by 100.

```
SELECT
    region,
    SUM(hydropower_and_pumped_storage + wind + solar) AS
total_renewable_energy,
    (
        SUM(hydropower_and_pumped_storage + wind + solar) /
SUM(net_generation)
    ) * 100 AS renewable_energy_percentage
FROM
    intel.energy_data
GROUP BY
    region
ORDER BY
    renewable_energy_percentage DESC;
```

- D. Which regions change from the top 3 when looking at total renewable energy vs percentage of renewable energy?

Northwest dominates both categories followed by Texas for the second place and Central in the third place when looking at **total renewable energy**. When we look at the **percentage of renewable energy**, Central is second place and California is third place.

– Task 2: Generating New Data by Energy Type

Intel would like to know how renewable energy and fossil fuels trend over time. In order to do this, you will first need to generate a new table using your SQL

knowledge and the `intel.energy_data` table before visualizing trends in Tableau Cloud.

- A.** Write a query that calculates the renewable energy generated for each row. Return only the `date`, `region`, and `energy_generated_mw` columns.

Note: `energy_generated_mw` is the alias for `hydropower_and_pumped_storage + wind + solar`.

```
SELECT
    date,
    region,
    (hydropower_and_pumped_storage + wind + solar) AS
    energy_generated_mw
FROM
    intel.energy_data;
```

After showing the result of the query to your manager, she tells you that she wants it to be clear that the `energy_generated_mw` column is referring to renewable energy types. She asks you to create a new column called `energy_type` that has the value 'renewable energy' for each row.

A colleague teaches you a simple method to do this. When writing your query, add an additional column after your select statement. Here is an example:

```
SELECT
    *, -- any relevant fields to the query
    'renewable energy' AS energy_type
FROM intel.energy_data
```

- B.** Modify your query from Part **A.** to include the `energy_type` column.

```
SELECT
    date,
    region,
    (hydropower_and_pumped_storage + wind + solar) AS
energy_generated_mw,
    'renewable energy' AS energy_type
FROM
    intel.energy_data;
```

- C.** Next, write a **new** query that calculates the fossil fuel energy generated for each row. As in Part **A.**, return only the `date`, `region`, and `energy_generated_mw` columns, where `energy_generated_mw` is now the alias for `all_petroleum_products + coal + natural_gas + nuclear + other_fuel_sources`.

```
SELECT
    date,
    region,
    (
        all_petroleum_products + coal + natural_gas + nuclear
    ) AS energy_generated_mw
FROM
    intel.energy_data;
```

- D.** Modify your query in Part **C.** to include the `energy_type` column. This column should have the value 'fossil fuel' for each row.

HINT: This is very similar to Part **B.**!

```
SELECT
    date,
    region,
    (all_petroleum_products + coal + natural_gas + nuclear)
AS energy_generated_mw,
    'fossil fuel' AS energy_type
```

```
FROM
    intel.energy_data;
```

- E. Your queries from Parts **B.** and **D.** should both have the columns `date`, `region`, `energy_generated`, and `energy_type`. Write one final query that **UNIONS** these two together.

```
SELECT
    date,
    region,
    (hydropower_and_pumped_storage + wind + solar) AS
energy_generated_mw,
    'renewable energy' AS energy_type
FROM
    intel.energy_data
UNION
SELECT
    date,
    region,
    (
        all_petroleum_products + coal + natural_gas + nuclear
    ) AS energy_generated_mw,
    'fossil fuel' AS energy_type
FROM
    intel.energy_data;
```


Task 3: Aggregating Power Plant Data

Intel has provided you with additional data in order to reach the best conclusion about the location of its next data center. In this task you will be working with two tables `intel.power_plants` and `intel.energy_by_power_plant`. You will need to join these tables before you can aggregate them to help the Intel team with their analysis.

- A. Join the `intel.power_plants` and `intel.energy_by_power_plant` data on the `plant_code`. This joined table will form the basis for the rest of the task.

If done correctly, your output will have 2,504 rows.

```
SELECT
  pp.plant_name,
  pp.plant_code,
  pp.region,
  pp.state,
  pp.primary_technology,
  ep.energy_type,
  ep.energy_generated_mw
FROM
  intel.power_plants pp
  JOIN intel.energy_by_plant ep ON pp.plant_code =
    ep.plant_code;
```

Note: It is recommended to use the **WITH** keyword for the remainder of this Task to simplify your queries. For a refresher, rewatch “ The **WITH** Keyword” in SkillBuilder 6.

- B. Write a query that returns the total number of **renewable energy** power plants for each region. Which region has the most renewable power plants?

```
SELECT
  pp.region,
  COUNT(DISTINCT pp.plant_code) AS total_renewable_plants
FROM
  intel.power_plants pp
  JOIN intel.energy_by_plant ep ON pp.plant_code =
    ep.plant_code
WHERE
```

```
    ep.energy_type = 'renewable_energy'
GROUP BY
    pp.region
ORDER BY
    total_renewable_plants DESC;
```

Midwest has the most renewable power plants , 234.

- C. Next, write a query that returns both the total number of power plants and the total energy generated, specifically from plants that use “Solar Photovoltaic” technology, grouped by each region.

```
SELECT
    pp.region,
    COUNT(DISTINCT pp.plant_code) AS total_solar_plants,
    SUM(ep.energy_generated_mw) AS total_energy_generated_mw
FROM
    intel.power_plants pp
    JOIN intel.energy_by_plant ep ON pp.plant_code =
    ep.plant_code
WHERE
    pp.primary_technology = 'Solar Photovoltaic'
GROUP BY
    pp.region
ORDER BY
    total_energy_generated_mw DESC;
```

- D. Modify your query in part C to only show regions having at least 50 power plants that use “Solar Photovoltaic” technology. What can you infer about the efficiency (or size) of the power plants in the Midwest region relative to the other regions in your output?

```
SELECT
    pp.region,
    COUNT(DISTINCT pp.plant_code) AS total_solar_plants,
    SUM(ep.energy_generated_mw) AS total_energy_generated_mw
FROM
    intel.power_plants pp
    JOIN intel.energy_by_plant ep ON pp.plant_code =
    ep.plant_code
WHERE
    pp.primary_technology = 'Solar Photovoltaic'
GROUP BY
    pp.region
HAVING
    COUNT(DISTINCT pp.plant_code) >= 50
ORDER BY
    total_energy_generated_mw DESC;
```

From the data, I can infer that the Midwest, with 71 Solar Photovoltaic plants, has a comparable number of plants to regions like Texas (57 plants) and California (59 plants). However, the total energy generated in the Midwest (4,907,305 MW) is much lower than in those regions, suggesting that the Solar Photovoltaic plants in the Midwest are either smaller or less efficient. This makes me think that the plants in the Midwest may have lower capacity or are using less efficient technology compared to the larger or more efficient plants in regions like Texas and California.

Note: There is more Tableau work up ahead! If you want to skip the LevelUp jump straight to **Task 4** below!

– LevelUp: Hourly Trends in Renewable Energy

Before moving on to your Tableau Visualizations, let's investigate how renewable energy generation fluctuates with the time of day.

- A.** Write a query that calculates the total **renewable** energy generated in each region for each hour of the day.

HINT: You'll need to use the `date_part` function to get the hour from the `time_at_end_of_hour` column. Your result should only have the values 0–23 for that new column.

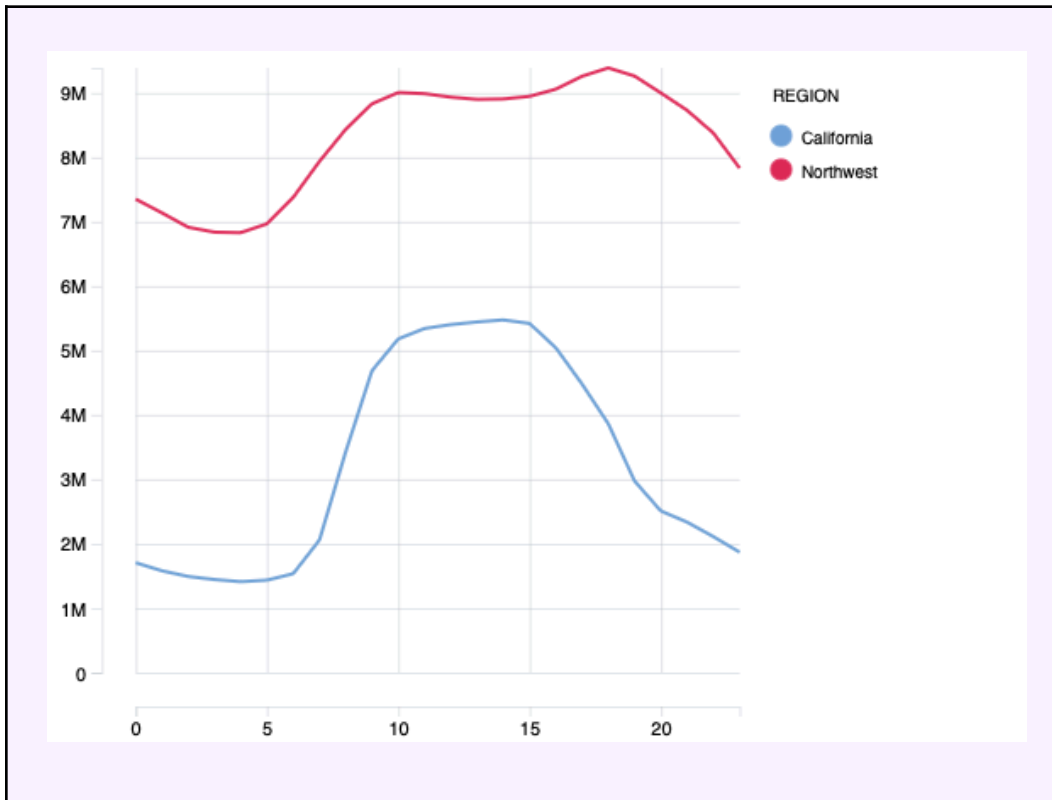
```
SELECT
    region,
    DATE_PART('hour', time_at_end_of_hour) AS hour_of_day,
    SUM(solar + wind + hydropower_and_pumped_storage) AS
total_renewable_energy_mw
FROM
    intel.energy_data
GROUP BY
    region,
    hour_of_day
ORDER BY
    region,
    hour_of_day;
```

- B.** Modify your query to filter to the 'California' and 'Northwest' regions only.

```
SELECT
    region,
    DATE_PART('hour', time_at_end_of_hour) AS hour_of_day,
    SUM(solar + wind + hydropower_and_pumped_storage) AS
total_renewable_energy_mw
FROM
    intel.energy_data
WHERE
```

```
region IN ('California', 'Northwest')
GROUP BY
  region,
  hour_of_day
ORDER BY
  region,
  hour_of_day;
```

- C. Use the built-in visualizer in the SQL app to plot a line graph of the energy generated for each hour of the day and colored by the region. If done correctly you should have two lines in your visualization.



- D. What can you say about the renewable energy generation between California (CAL) and the Pacific Northwest (NW)?

California shows a steep increase in renewable energy generation starting from around 7 AM, peaking between 10 AM and 3 PM, then gradually declining towards the evening. This pattern aligns with the reliance on solar energy, where peak generation occurs during midday when sunlight is strongest.

Northwest, on the other hand, has a more consistent and higher level of renewable energy generation throughout the day, with a much flatter curve. This suggests that the NW region may rely more on renewable sources like wind or hydropower, which are less dependent on sunlight and can generate energy more steadily across different hours.

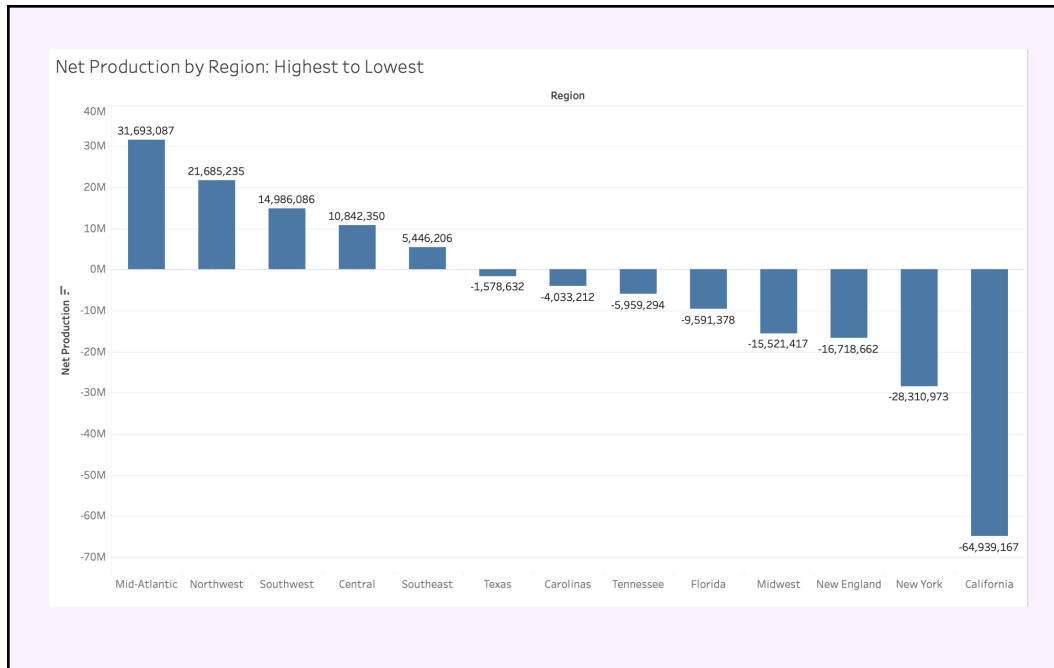
In summary, the Pacific Northwest appears to have more stable and higher renewable energy production across the day, while California's renewable energy generation is more concentrated during daylight hours, likely due to solar energy.

– Task 4: Visualizing and Analyzing Using Tableau

Continue to post your answers in the provided boxes: **purple boxes** for your visualizations, and **blue boxes** for text-based answers.

- A. On the “Net Production” sheet, create a bar chart of net production, by region. Sort the chart in *descending* order, from tallest to smallest.

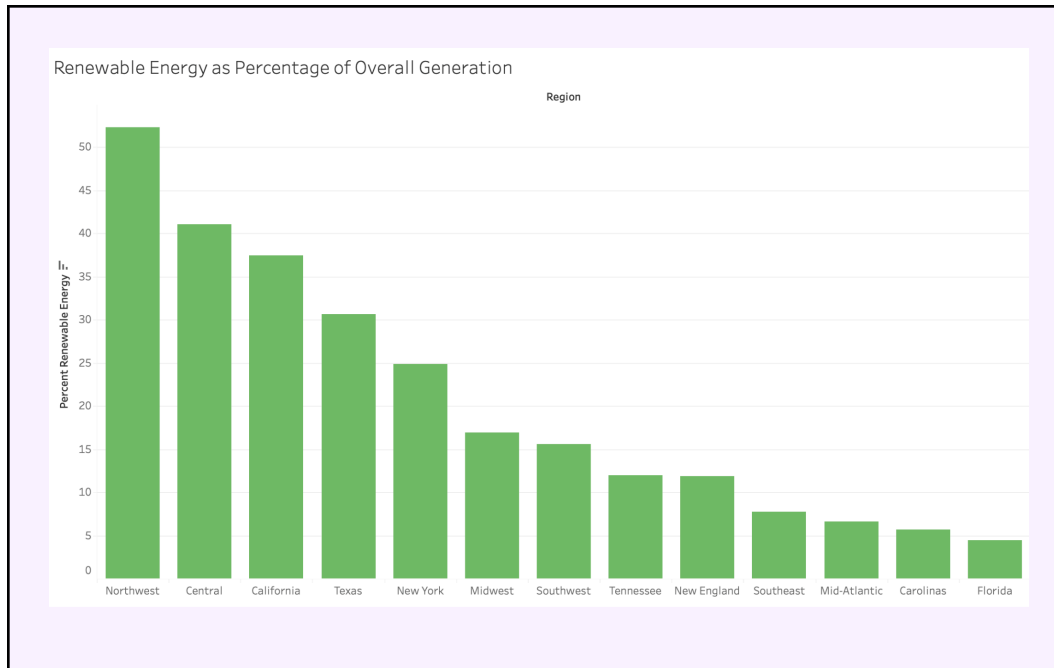
The net energy produced is calculated by subtracting the total energy demand from the total energy generation. This is already created in the field called **Net Production**.



B. Next, on the “Renewable Energy” sheet, create a bar chart illustrating which regions generate the greatest percentage of renewable energy.

HINT: In Tableau, you have a field called `Percent Renewable Energy`

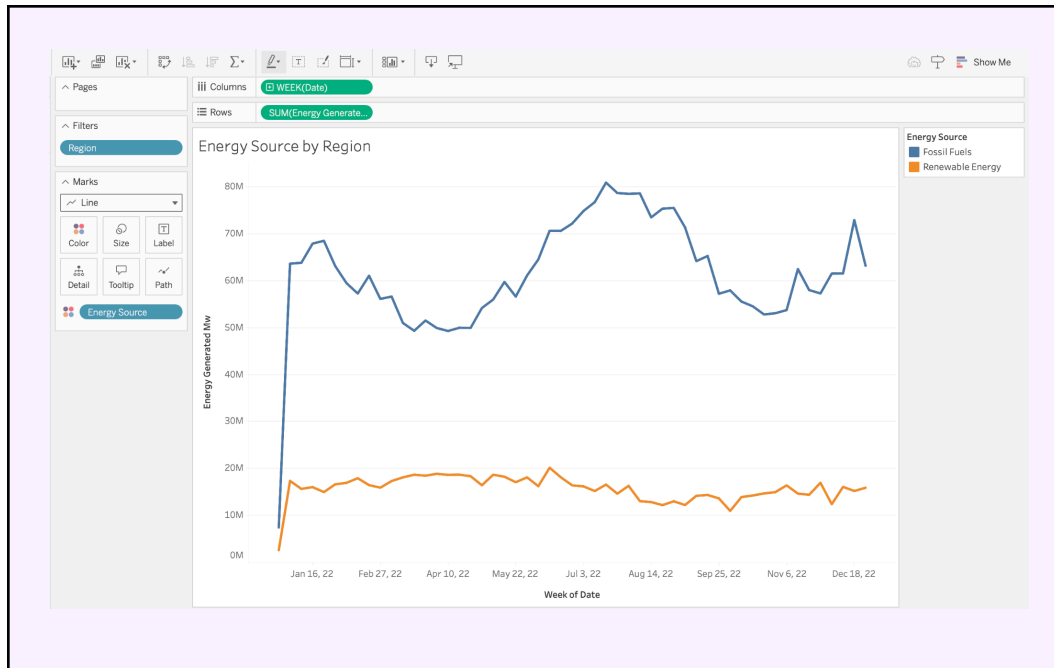
Create a bar chart in descending order of regions with the most renewable energy percentage.



- C. On the “Energy Source by Region” sheet, create a line chart of the energy generated for each energy source (fossil fuels & renewable energy) at the weekly date level. Add a filter for the region to your chart.

For this chart, you will use the `energy_by_source` dataset loaded into your Tableau workbook.

Remember to include your pills and filters in the screenshot

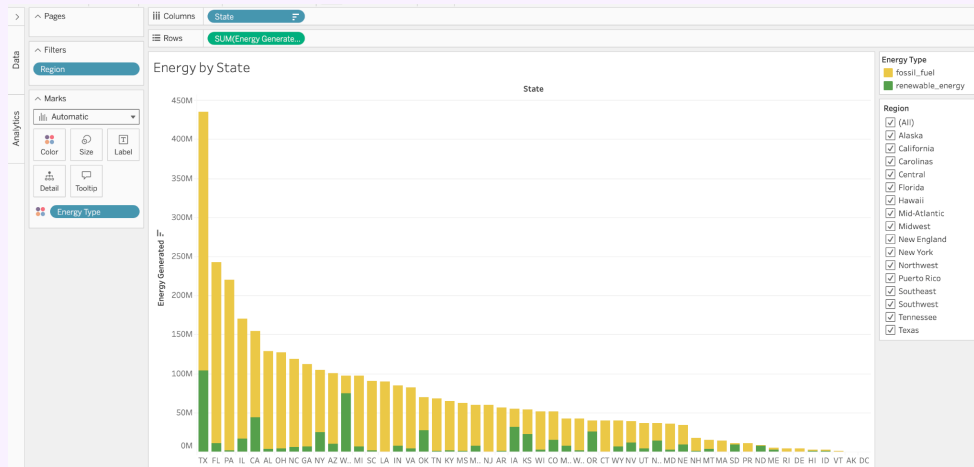


- D.** On the “Energy by State” sheet create a bar chart of the total energy generated by each state and energy type. Color the bars by energy type. Include a region filter in your chart to reduce the amount of bars shown.

For this chart you will use the `power_plant_energy` dataset that you created. You can select the data source in the upper left hand column in Tableau.

(paste your visualization screenshot here)

Remember to include your pills and filters in the screenshot

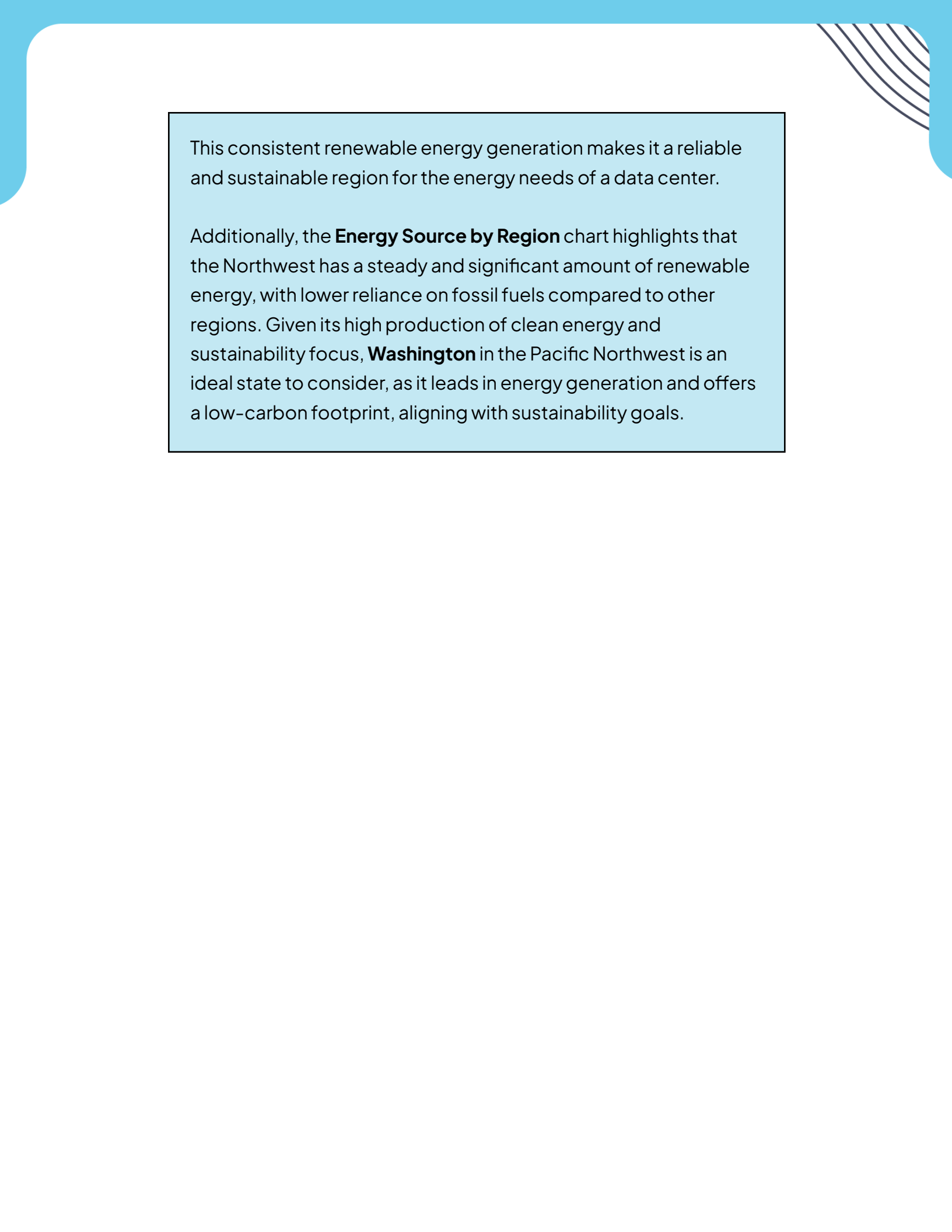


– Task 5: Communicating Results

Your manager wants you to share the visualizations you created in Task 3 with the Sustainability team for visibility. She has created a dashboard with your visualizations (see the “Dashboard” sheet in Tableau) and has asked you to write a short paragraph explaining which region you recommend that the next data center be built.

- A. In 1–2 paragraphs, summarize what can be gleaned from your visualizations. What **region** and **state** do you think is best and why?

From the dashboard, we can observe that the **Pacific Northwest** emerges as a strong contender for building the next data center. The **Net Production by Region** chart shows that the Northwest is one of the top regions for net energy production, particularly in renewable energy, as indicated by the **Renewable Energy as a Percentage of Overall Generation** chart, where it ranks highest.



This consistent renewable energy generation makes it a reliable and sustainable region for the energy needs of a data center.

Additionally, the **Energy Source by Region** chart highlights that the Northwest has a steady and significant amount of renewable energy, with lower reliance on fossil fuels compared to other regions. Given its high production of clean energy and sustainability focus, **Washington** in the Pacific Northwest is an ideal state to consider, as it leads in energy generation and offers a low-carbon footprint, aligning with sustainability goals.