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.

— 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?

```
SELECT
  region,
```

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

The Mid-Atlantic region has the highest positive total energy, totalling 31,693,087 MW.

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?

```
SELECT
  region,
  SUM(hydropower_and_pumped_storage + solar + wind) AS
total_renewable_energy_generated
FROM
  intel.energy_data
GROUP BY
  region
```

```
ORDER BY
  total_renewable_energy_generated DESC
```

The top two regions in renewable energy production are the Northwest and Texas regions, producing 199,266,574 and 131,367,234 MW respectively.

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

```
SELECT
  region,
  ROUND(SUM(hydropower_and_pumped_storage + solar + wind) /
SUM(net_generation) * 100,2) AS percent_renewable_energy
FROM
  intel.energy_data
GROUP BY
  region
ORDER BY
  percent_renewable_energy DESC
```

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

Northwest, Texas, and Central regions were top 3 when looking at total renewable energy, but when looking at the percentage of renewable energy, the Northwest, Central, and California regions are top 3, and Texas falls in fourth in terms of percentage of

renewable energy compared to its position as second in total renewable energy.

— **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,
  SUM(hydropower_and_pumped_storage + solar + wind) AS
energy_generated_mw
FROM
  intel.energy_data
GROUP BY region,date
```

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

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

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,
  SUM(all_petroleum_products + coal + natural_gas + nuclear +
other_fuel_sources) AS energy_generated_mw
FROM
  intel.energy_data
GROUP BY region,date
```

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

```
SELECT
  date,
  region,
  SUM(all_petroleum_products + coal + natural_gas + nuclear +
other_fuel_sources) AS energy_generated_mw,
  'fossil fuel' AS energy_type
FROM
  intel.energy_data
GROUP BY region,date
```

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 `UNION`s these two together.

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

```
  intel.energy_data
GROUP BY region,date
```

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.

```
SELECT
    e.*,
    p.region,
    p.state,
    p.fuel_types,
    p.primary_technology
  FROM
    intel.energy_by_plant AS e
    INNER JOIN intel.power_plants AS p ON e.plant_code =
p.plant_code
```

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?

```
WITH plant_table AS (
  SELECT
    e.*,
    p.region,
    p.state,
    p.fuel_types,
    p.primary_technology
  FROM
    intel.energy_by_plant AS e
    INNER JOIN intel.power_plants AS p ON e.plant_code =
p.plant_code

)

SELECT region, COUNT(energy_type) AS
num_renewable_power_plants
FROM plant_table
WHERE energy_type ILIKE 'renewable_energy'
GROUP BY region
ORDER BY num_renewable_power_plants DESC
```

The Midwest has 234 renewable power plants, which is the most of any region in the dataset.

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.

```
WITH plant_table AS (
  SELECT
```

```
    e.*,
    p.region,
    p.state,
    p.fuel_types,
    p.primary_technology
  FROM
    intel.energy_by_plant AS e
    INNER JOIN intel.power_plants AS p ON e.plant_code =
p.plant_code
)
SELECT
  region,
  COUNT(plant_code) AS num_plants,
  SUM(energy_generated_mw) AS total_energy_generated_mw
FROM
  plant_table
WHERE
  primary_technology ILIKE 'Solar Photovoltaic'
GROUP BY region
```

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?

```
WITH plant_table AS (
  SELECT
    e.*,
    p.region,
    p.state,
    p.fuel_types,
    p.primary_technology
  FROM
    intel.energy_by_plant AS e
```

```
    INNER JOIN intel.power_plants AS p ON e.plant_code =
p.plant_code
)
SELECT
  region,
  COUNT(plant_code) AS num_plants,
  SUM(energy_generated_mw) AS total_energy_generated_mw
FROM
  plant_table
WHERE
  primary_technology ILIKE 'Solar Photovoltaic'
GROUP BY region
HAVING COUNT(plant_code) >= 50
```

Although the Midwest has the third most power plants that use primarily Solar Photovoltaic technology, it has the lowest output of energy in comparison to other regions, which suggests that the power plants are inefficient or the Midwest has smaller sized power plants that don't output as much energy.

– **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.

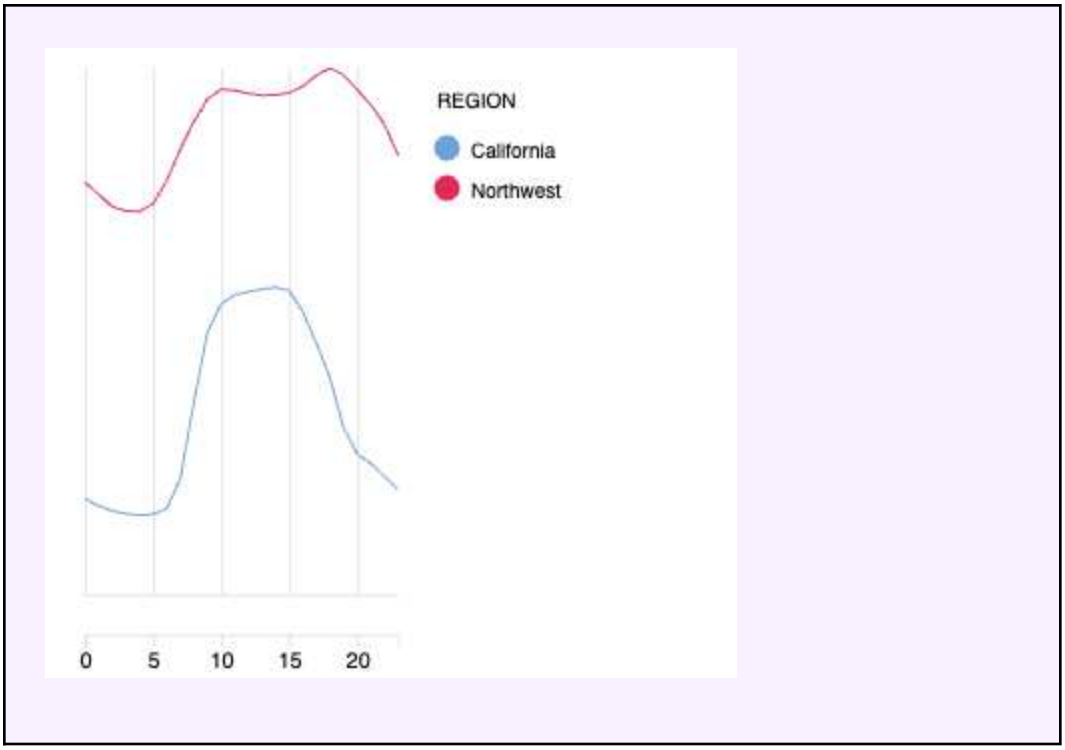
A. Write a query that calculates the total **renewable** energy generated in each region for each hour of the day.

```
SELECT
  date_part('hour',time_at_end_of_hour) AS hour,
  region,
  SUM(hydropower_and_pumped_storage + solar + wind) AS
total_renewable_energy_generated
FROM
  intel.energy_data
GROUP BY
hour,region
```

B. Modify your query to filter to the ‘California’ and ‘Northwest’ regions only.

```
SELECT
  date_part('hour',time_at_end_of_hour) AS hour,
  region,
  SUM(hydropower_and_pumped_storage + solar + wind) AS
total_renewable_energy_generated
FROM
  intel.energy_data
WHERE region ILIKE 'California' OR region ILIKE 'Northwest'
GROUP BY
hour,region
```

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)?

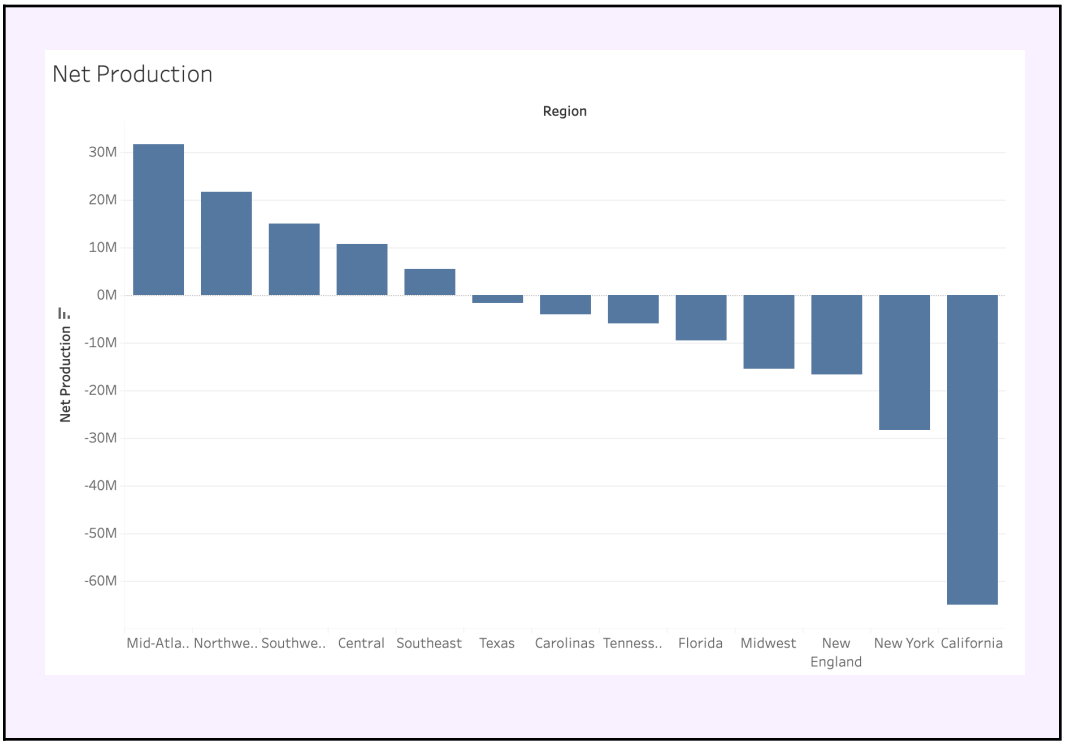
> The Northwest generally generates more renewable energy than California and both regions generate most of their renewable energy during daytime hours.

— **Task 4:** Visualizing and Analyzing Using Tableau

Phew! Now that you've gotten the queries out of the way, you're ready to dive into investigating the best regions for Intel to put its next data center. The remaining Tasks will be completed in Tableau, and will focus on visualizing and analyzing your results.

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.

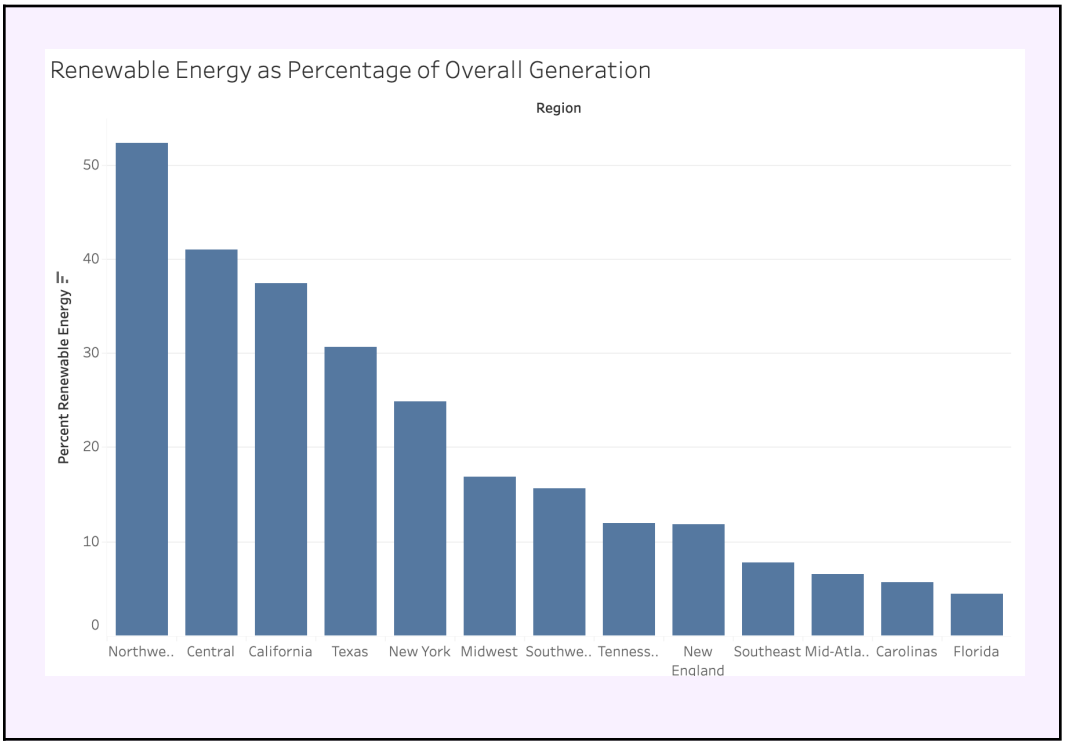
The net energy produced is calculated by subtracting the total energy demand from the total energy generation. This is already created in the field called `Net Production`.



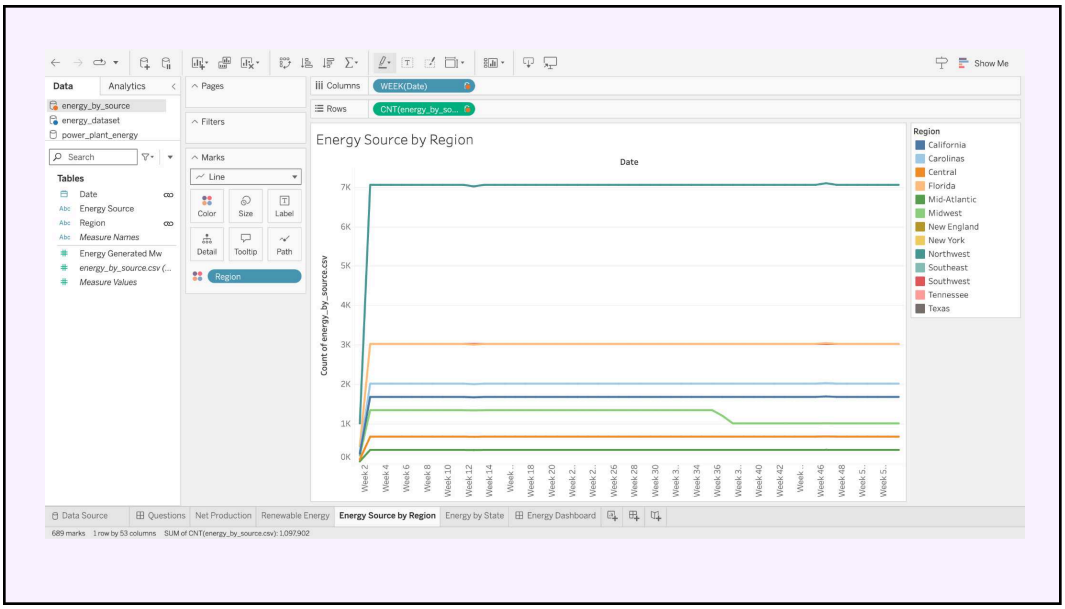
B. Next, on the “Renewable Energy” sheet, create a bar chart illustrating which regions generate the greatest percentage of renewable energy.

.

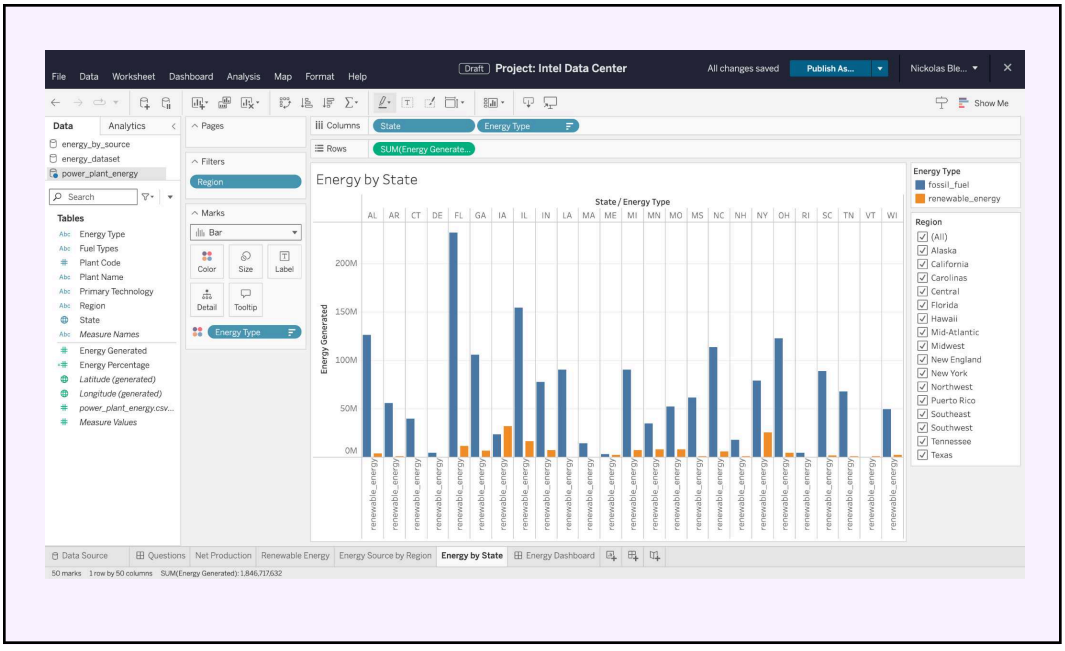
Create a bar chart in descending order of regions with the most renewable energy percentage.



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.



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.



– **Task 5:** Communicating Results

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

Based on the dashboard, We can see that the Mid-Atlantic, Northwestern, Southwestern, Central, and Southeastern regions all produce a surplus of energy, and are therefore more likely to provide energy at cheaper prices. However, when taking renewable energy into account, the Northwest and Central regions have the highest percentages of renewable energy out of their total energy output.Renewable energy makes up 52.25% and 41.01% of the total energy produced in the Northwest and Central regions respectively, while in the Mid-Atlantic it's 6.58%, in the Southwest it's 15.57%, and in the Southeast it's 7.75%. Thus, after narrowing it down to states in these two regions, we must now look at these states individually. From the Energy by State chart, we can see that Washington produces the most amount of renewable energy and also produces the most amount of energy as a whole out of every state in the two regions. Thus, I believe Washington is the best state to build an Intel data center. It's located in a region that produces a surplus of energy, so energy prices may be cheaper, and renewable energy makes up more than 50% of its total energy output, fitting in with Intel's green initiative.