Intel Data Center

intel®

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. When you're done, export your document as a pdf file and submit it on the Milestone page – see instructions for creating a PDF at the end of the Milestone.

RESOURCES: If you need hints on the Milestone or are feeling stuck, there are multiple ways of getting help. Attend Drop-In Hours to work on these problems with your peers, or reach out to the HelpHub if you have questions. Good luck!

SQL App: **Here’s that link** to our specialized SQL app, where you’ll write your SQL queries and interact with the data.

— 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) AS NET_GEN,
  SUM(demand) AS dem,
  SUM(net_generation) - SUM(demand) AS tot_energy
FROM intel.energy_data
GROUP BY region
ORDER BY tot_energy DESC
```

In certain region the demand is higher then the net generation, means that they need to buy energy from other region to fill the gap.

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 Renewal_energy
FROM intel.energy_data
GROUP BY region
ORDER BY Renewal_energy DESC
```

The top Two are:
Northwest: 199 266 574 MegaWatts
Texas: 131 367 234 MegaWatts

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 Renewal_energy,
  SUM(net_generation) AS net,
  SUM(hydropower_and_pumped_storage + wind +
solar) / SUM(net_generation) * 100 AS
Renewal_energy_by_percent
FROM
  intel.energy_data
GROUP BY
  region
ORDER BY
  Renewal_energy_by_percent DESC
```

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

Texas is not anymore in the top 3.
Percentage of renewable energy of the top 3 are:

Northwest: 52.252%
Central: 41.013%
California: 37.453

— **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 + wind +
solar) AS energy_generated_mw
FROM
  intel.energy_data
GROUP BY
  region, date
ORDER BY
  energy_generated_mw DESC
```

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

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
ORDER BY
  energy_generated_mw DESC
```

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,
  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
```

```
ORDER BY
  energy_generated_mw DESC
```

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 + wind +
solar) AS energy_generated_mw,
  'renewable energy' AS energy_type
FROM
  intel.energy_data
GROUP BY
  region, date
ORDER BY
  energy_generated_mw DESC)

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

UNION ALL

(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
ORDER BY
  energy_generated_mw DESC)
```

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
  *
FROM
  intel.power_plants AS pow
  JOIN intel.energy_by_plant AS en ON
en.plant_code = pow.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?

```
with power_energy as (
  SELECT
```

```
    *
  FROM
    intel.power_plants AS pow
    JOIN intel.energy_by_plant AS en ON
en.plant_code = pow.plant_code
  WHERE
    en.energy_type = 'renewable_energy'
)
SELECT
  region,
  COUNT(*) AS num_plant
FROM
  power_energy
GROUP BY
  region
ORDER BY
  num_plant DESC
```

MIDWEST has the most renewable energy

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 power_energy as (
  SELECT
    *
  FROM
    intel.power_plants AS pow
    JOIN intel.energy_by_plant AS en ON
en.plant_code = pow.plant_code
    WHERE primary_technology = 'Solar
Photovoltaic'

)
SELECT
    region,
    COUNT(*) AS num_plant,
    SUM(energy_generated_mw) as eg_mw
FROM power_energy
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 power_energy as (
  SELECT
    *
  FROM
    intel.power_plants AS pow
    JOIN intel.energy_by_plant AS en ON
en.plant_code = pow.plant_code
    WHERE primary_technology = 'Solar
Photovoltaic'

)
SELECT
    region,
    COUNT(*) AS num_plant,
    SUM(energy_generated_mw) as eg_mw
FROM power_energy
GROUP BY region
```

```
HAVING COUNT(*) >=50
```

Midwest region has the lowest total energy generated despite having 71 Solar photovoltaic plant.

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.

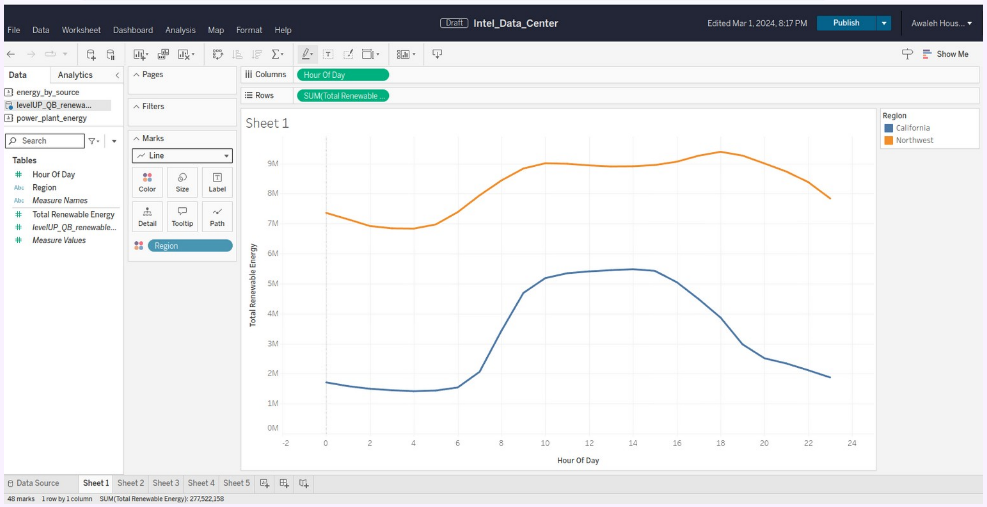
HINT: You’ll need to use the `date_part` function to get the hour from the `time_at_end_of_hour` column. Your result should only have the values 0-23 for that new column.

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

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(hydropower_and_pumped_storage + wind +
solar) AS total_renewable_energy
FROM
  intel.energy_data
WHERE
  region = 'California' OR region= 'Northwest'
GROUP BY
  region,
  date_part('hour', time_at_end_of_hour)
ORDER BY
  region,
  date_part('hour', time_at_end_of_hour);
```

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.



Columns: Hour Of Day

Rows: SUM(Total Renewable ...

48 marks 1 row by 1 column SUM(Total Renewable Energy): 277,522,158

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

The renewable energy start increase around 6:00am and start decreasing after 18:00 or (6:00pm)

— **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. **Click this link to navigate to the workbook you'll use to complete the remainder of this Project.**

Once you've published your Tableau Workbook, paste the Share Link in the box below.

Note: Your share link must begin with:

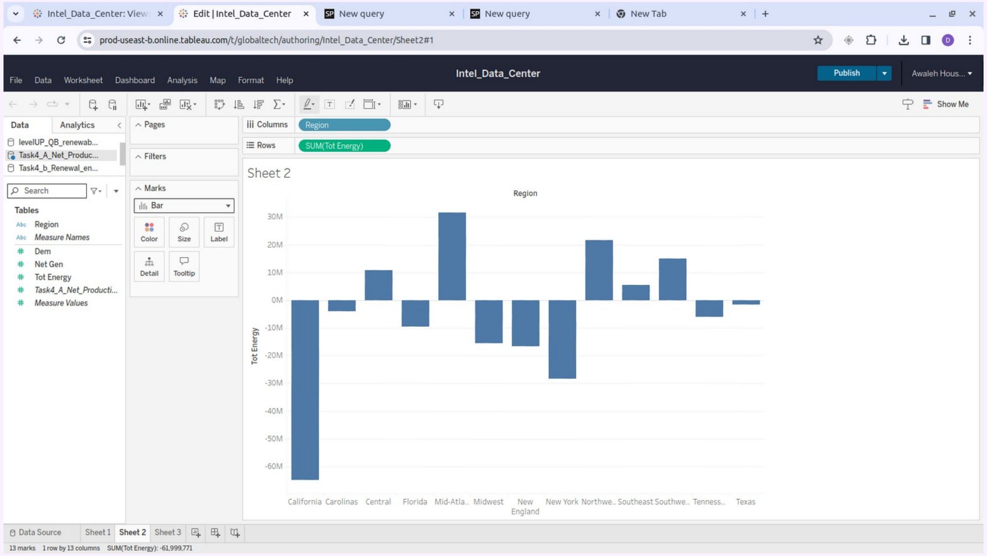
`https://prod-useast-b.online.tableau.com/#/site/globaltech/workbooks/...`

https://prod-useast-b.online.tableau.com/#/site/globaltech/workbooks/1318456?:origin=card_share_link

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`.

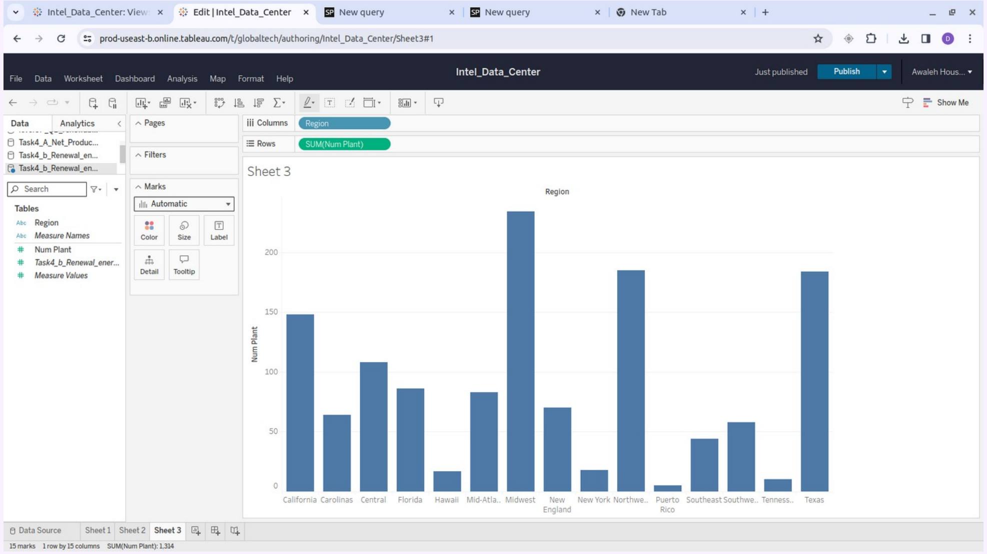


B. Next, on the “Renewable Energy” sheet, create a bar chart illustrating which regions generate the greatest percentage of renewable energy.

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

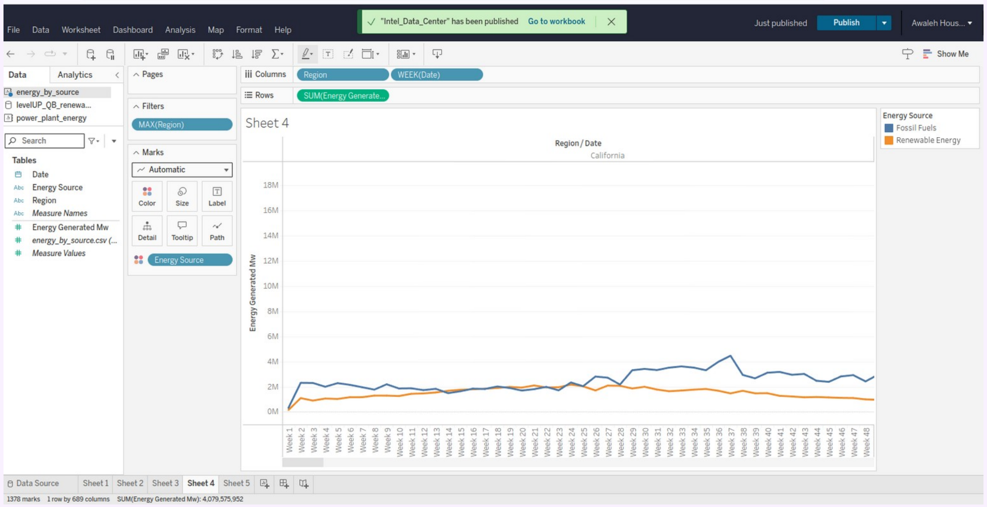
.

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.



"Intel_Data_Center" has been published Go to workbook

Columns: Region, WEEK(Date)

Rows: SUM(Energy Generate...

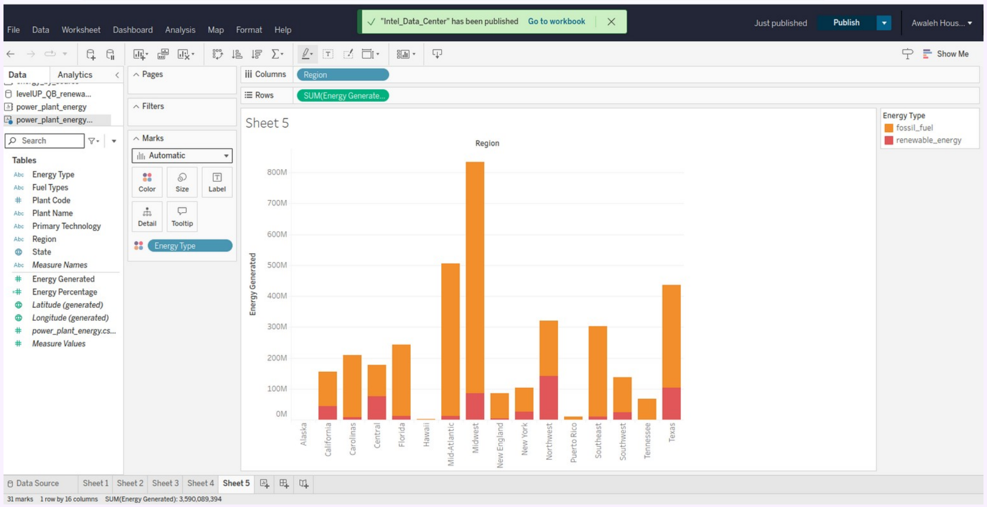
Filters: MAX(Region)

Marks: Automatic; Color: Energy Source

1378 marks 1 row by 689 columns SUM(Energy Generated Mw): 4,079,575,952

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.



✓ "Intel_Data_Center" has been published

31 marks 1 row by 16 columns SUM(Energy Generated): 3,590,089,394

— **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?

The **Northeast** is the best location to build the next data center. The northeast produce high percentage of renewable energy compare to other region. It is a land where natural resources such us wind and solar are abundant. It is friendly environment and contribute to reduce carbon emission. **Massachusetts** boasts ample sunlight, making it an ideal location for installing a data center.

That’s it! Submit your final project for evaluation, and go celebrate your achievement! You just completed a rich, complex data analysis project representing real-world level work. You’ve gained some impressive skills! Well done, and never stop learning 😀

— Submission

Great work completing your Portfolio Project!! To submit your completed project file, you will need to download / export this document as a PDF and then upload it to the Milestone submission page. You can find the option to download as a PDF from the File menu in the upper-left corner of the Google Doc interface. Triple check the link to your Tableau workbook.

Evaluation Rubric

Task	Task title	Max points
1	Energy Generation	50
2	Generating a New Table by Energy Type	55
3	Aggregating Power Plant Data	35
4	Visualizing Using Tableau	40
5	Communicating Results	20
	TOTAL POINTS:	200
	LevelUp	
1	Hourly Trends in Renewable Energy	20