CS653

Final Project Report

The Impact of

Climate Change

on Winter Tourism

--The Future of Winter Tourism in a Warming World –

By

Teevit Lertchaturaporn - 6509035082
Rattakorn Keatprakob - 6509035165
Ratchapol Marmongkol - 6509035033
Prerapong Ramunudom - 6509035223
Worrapon Thongmook - 6509035025
Kunanon Sukjaruen - 6509035207

Table of Contents

Part I Introduction & Objectives..... 3
Part II Dataset..... 5
Part III 5Vs of Data..... 7
Part IVArchitecture of Data Pipeline..... 8
Part V Data pipeline Implementation..... 10
Part VI Result, Conclusion and Discussion..... 13
Part VII Reflection & Suggestion for Further Studies..... 15
Part VIII Project Resource..... 16
Reference..... 17

Part I
Introduction & Objectives

Global warming and climate change are having a significant impact on the world's weather patterns. It is caused by the release of greenhouse gasses into the atmosphere, which trap heat and cause the planet to warm. Some of the common effects of climate change include:

- **Rising sea levels:** As the planet warms, the ice caps and glaciers are melting, which is causing sea levels to rise. This is a threat to coastal communities around the world.
- **More extreme weather events:** Climate change is causing more extreme weather events, such as hurricanes, floods, and droughts. These events can cause widespread damage and loss of life.
- **Changes in plant and animal life:** Climate change is causing changes in plant and animal life. Some species are moving to new areas in search of cooler temperatures, while others are becoming extinct.
- **Changes in agriculture:** Climate change is making it more difficult to grow crops in some areas. This is a threat to food security.

The effects of climate change on weather patterns are complex and vary from region to region. However, some general trends can be observed. In general, temperatures are rising worldwide, and this is leading to changes in precipitation patterns. In some areas, this is leading to more droughts, while in others it is leading to more flooding. It is also causing changes in the frequency and intensity of extreme weather events, such as heat waves, storms, and floods. These events can have a devastating impact on people and the environment [1].

The impact of climate change on snow in the U.S. is already being felt. The amount of snow that falls has decreased significantly since the 1970s [2]. This could lead to shorter ski seasons and less snow for winter recreation. Climate change is also causing changes in the timing of snowfall, which is making it more difficult to plan for winter activities. The correlation between temperature and revenue from the ski industry may have reasonable trends. As temperatures rise, ski resorts may have fewer visitors and lower revenue. This is because skiers are less likely to visit resorts when the snow is not reliable.

The news media has also been reporting on the effects of climate change on the ski industry [3]. In recent years, there have been numerous articles and reports about how climate change is making it harder for ski resorts to operate.

Therefore, understanding how global warming will impact tourism in various locations can help us plan and adapt tourism strategies in a timely and sustainable manner. Big data analysis can help tourism businesses to adapt to the challenges posed by global warming. By analyzing data on tourism trends, weather patterns, and environmental conditions, businesses can identify potential risks and opportunities. This information can then be used to develop strategies that will help businesses to mitigate risks and capitalize on opportunities.

The purpose of this study is to assess the impact of climate change on the U.S. land temperature and average snow pattern. The study will examine the following:

1. The use of Amazon Web Services (AWS) to utilize big data analytics to tackle the global warming and climate change problem.
2. The impact of the U.S. land temperature on snow coverage area.
3. The analytics of time series data on how the U.S. tourism industry needs to adapt the strategies.

Part II

Dataset

The study uses the “NOAA U.S. Climate Gridded Dataset (NClimGrid)” to make an analysis of the changing of temperature in all states over the U.S. together with “Monthly Area of Snow Extent” to make a descriptive analysis of snow coverage area in North America Region.

EpiNOAA - NOAA U.S. Climate Gridded Dataset (NClimGrid)

The first dataset, NOAA U.S. Climate Gridded Dataset (NClimGrid) [4], consists of four climate variables derived from the GHCN-D dataset: maximum temperature, minimum temperature, average temperature and precipitation. Each file provides monthly values in a 5x5 lat/lon grid for the Continental United States. Data is available from 1895 to the present. For this study we use a derived version of the dataset called “EpiNOAA”. EpiNOAA is an analysis ready dataset that consists of a daily time-series of nClimGrid measures (maximum temperature, minimum temperature, average temperature, and precipitation) at the county scale. Each file provides daily values for the Continental United States. Data is available from 1951 to the present. Daily data are updated every 3 days with a preliminary data file. In this dataset, there are data fields as in **Table 1**.

Table 1. Data fields of EpiNOAA - NOAA U.S. Climate Gridded Dataset

Field	**Description and format**
date	The date of the weather observation (format : mm/dd/yyyy)
year	The year of the weather observation (format : yyyy)
month	The month of the weather observation (format : mm)
day	The day of the month of the weather (format : dd)
state	The state in which the weather observation was taken (format : XX two abbreviations characters)
county	The sub-area of the state in which the weather observation was taken
region_code	The code which indicate state-county
prcp	The amount of precipitation that fell on the day of the weather observation in millimeters (mm)
tavg	The average temperature for the day of the weather observation in degree celsius (^{O}C)
tmin	The minimum temperature for the day of the weather observation in degree celsius (^{O}C)
tmax	The maximum temperature for the day of the weather observation in degree celsius (^{O}C)

Monthly Area of Snow Extent

The second dataset, Monthly Area of Snow Extent [2], is from the Rutgers Global Snow Lab. It shows the monthly area of snow cover in the Northern Hemisphere. The data is available from 1967 to the present. The data is presented in a table which the fields are as in **Table 2**.

Table 2. Data fields of Monthly Area of Snow Extent dataset

Field	**Description and format**
Year	The year of the snow cover observation
Month	The month of the snow cover observation
N.Hemisphere	The area of snow cover in the Northern Hemisphere in million square kilometers
Eurasia	The area of snow cover in Eurasia in million square kilometers
N.America	The area of snow cover in North America in million square kilometers
N.America (no Greenland)	The area of snow cover in North America excluding Greenland in million square kilometers

Both datasets are fed to the system by using our constructed data pipeline ingestion, storage, analysis, and visualization to find the answer for the study.

Part III

5Vs of Data

Volume

The volume of data refers to the amount of data that is collected and stored. Big data can be terabytes, petabytes, or even exabytes in size. In the case of the NOAA U.S. Climate Gridded Dataset (NClimGrid), contains a significant volume of data as it covers the entire county in every state in the U.S. and includes temporal data since 1951 to present.

Velocity

The velocity of data refers to the speed at which data is generated and processed. Big data can be generated in real time, or it can be collected over a period of time. For NClimGrid, the data may exhibit varying rates of change over time, such as temperature fluctuations or rainfall patterns within short intervals.

Variety

The variety of data refers to the different types of data that are collected. Big data can include structured data, unstructured data, and semi-structured data. The datasets in this study come from two sources: the U.S. National Oceanic and Atmospheric Administration (NOAA) and the Rutgers University Snow Lab. The NOAA dataset includes county-level temperature and precipitation data, while the Rutgers dataset includes area of snow coverage data. Both datasets contain numerical and categorical data. Therefore, we must consider all layers of data analysis when working with these datasets.

Veracity

The veracity of data refers to the accuracy and reliability of the data. Big data can be noisy and contain errors. NClimGrid has undergone certification and validation by NOAA, a reputable government agency specializing in climatology. However, ensuring data veracity is essential through verification and assessment for reliable analysis.

Value

The value of data refers to the insights that can be gained from the data. Big data can be used to make better decisions, improve products and services, and identify new opportunities. NClimGrid holds significant value, enabling environmental planning, agricultural decision-making, public dissemination of weather information, and facilitating climate-related research within the United States. The value derived from such analyses can enhance planning and decision-making processes by providing more confidence in utilizing weather data. Where an area of snow coverage dataset can be used to identify a correlation of snow area and the terrestrial temperature.

Part IV

Architecture of Data Pipeline

Our study proposes a data pipeline architecture that is guided by the six pillars of the AWS well-architected framework. **Figure 1** shows the proposed data pipeline.

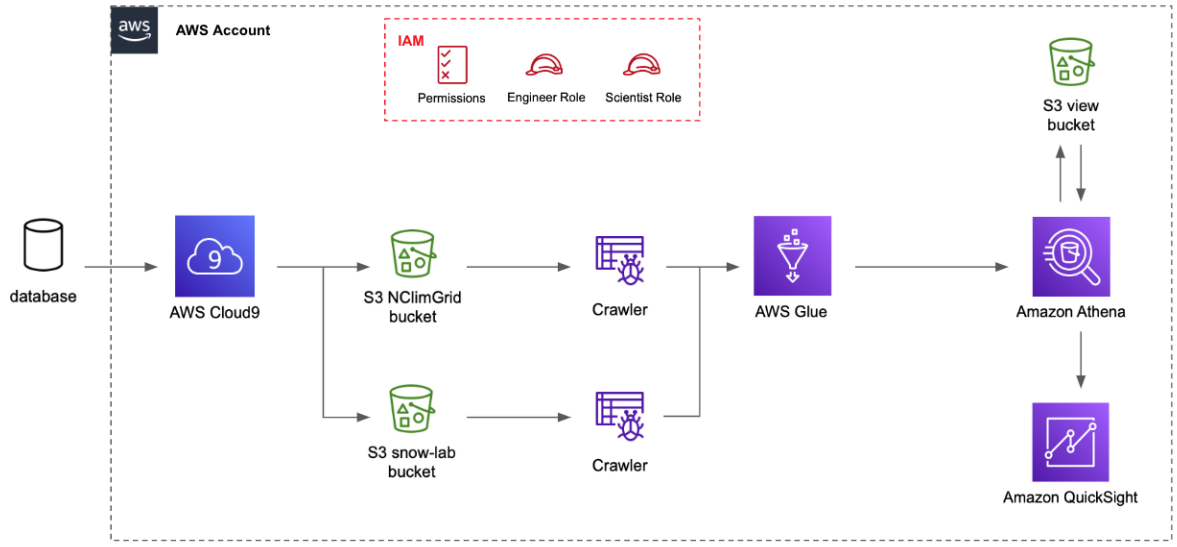


Figure 1. The proposed data pipeline. Beginning with the AWS Cloud9 for running scripts to get datasets from the outside of an AWS account then load into buckets. Then create crawlers to get the schemas of datasets and create databases and tables. After that, Amazon Athena can analyze data by querying data from the datasets and create views that will be stored in the view bucket. Finally, Amazon QuickSight can use the results in the view bucket to visualize the information.

The data pipeline is designed by considering the six pillars of the AWS Well-Architected Framework, as follows:

Operational excellence

The designed data pipeline can effectively achieve the analysis objectives. It minimizes operation costs through good cost optimization. It also requires less time for analysis as the results are stored in a bucket for visualization.

Reliability

AWS Glue is used to automatically discover the schema and metadata of data. In addition to that, a Data Engineer can also make further adjustments or modifications to the schema. This provides flexibility in handling data transformations and allows for adaptable changes to the incoming data.

Performance efficiency

Choosing to use Parquet file format, which is a column-based storage, allows for faster data access when retrieving data by columns. Additionally, Parquet files have a small file size, resulting in space-saving storage.

Security

During the work process, we follow the “least privilege” principles, an IAM role named “Engineer role” and “Scientist role” is assigned. This role effectively controls access, allowing limited usage of specific services such as S3, Amazon Athena, Amazon Quicksight, and AWS Glue. For the engineer role responsible for building the data pipeline, full access and utilization of all services are granted. However, for the scientist role, they are restricted to only viewing data within all buckets, querying data using Amazon Athena, and creating visualizations in Amazon Quicksight.

Cost optimization

Some portions of the data, originally in CSV files, are converted to Parquet format to reduce the storage size in S3. Athena is utilized to create views, which are then stored in S3 for visualization purposes in Amazon QuickSight. Storing the data as views helps reduce the amount of data being imported into Amazon QuickSight.

Sustainability

The project currently utilizes data only from the USA. However, in the future, if it becomes possible to incorporate temperature data from all countries around the world, the analysis coverage will significantly increase. This would enable analyzing the temperature of the entire globe consistently.

Part V

Data pipeline Implementation

1. Ingestion

We ingest 2 datasets into S3 buckets by AWS Cloud9 terminal: EpiNOAA - NOAA U.S. Climate Gridded Dataset (NClimGrid) and Monthly Area of Snow Extent. To import the datasets into our environment, we create an AWS Cloud9 environment and S3 buckets for storage.

1.1 Ingesting EpiNOAA - NOAA U.S. Climate Gridded Dataset (NClimGrid)

The dataset can be accessed with *https://registry.opendata.aws/noaa-nclim grid/*. After choosing *Browse Bucket* in the webpage, we find that the target dataset is stored in the path *epinoaa/parquet/*. Although the dataset is available in CSV format, we choose PARQUET format for storage efficiency optimization since it is much smaller in size than CSV.

In Cloud9 terminal we make loop commands to iterate the file's name, which contains year and month, for example “201809-cty-scaled.parquet” stores data collected in September 2018. 2 commands following downloads the data from year 1951-2021 into our EC2 instance.

(1) *for year in `seq 1951 2021`; do for month in `seq 1 9`;*
do wget "https://noaa-nclimgrid-daily-pds.s3.amazonaws.com/EpiNOAA/parquet/
${year}0${month}-cty-scaled.parquet"; done; done

(2) for year in `seq 1951 2021`; do for month in `seq 10 12`;
do wget "https://noaa-nclimgrid-daily-pds.s3.amazonaws.com/EpiNOAA/parquet/
${year}${month}-cty-scaled.parquet"; done; done

Afterwards, we run 2 following commands to upload them to our S3 bucket.

(1) *for year in `seq 1951 2021`; do for month in `seq 1 9`; do aws s3 cp*
"${year}0${month}-cty-scaled.parquet" s3://county-weather; done; done

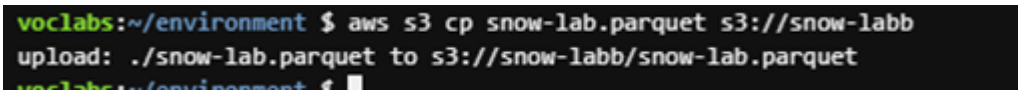
(2) *for year in `seq 1951 2021`; do for month in `seq 10 12`; do aws s3 cp*
"${year}${month}-cty-scaled.parquet" s3://county-weather; done; done

1.2 Ingesting Monthly Area of Snow Extent

We ingest this dataset with its presigned URL by Cloud9 terminal as follows

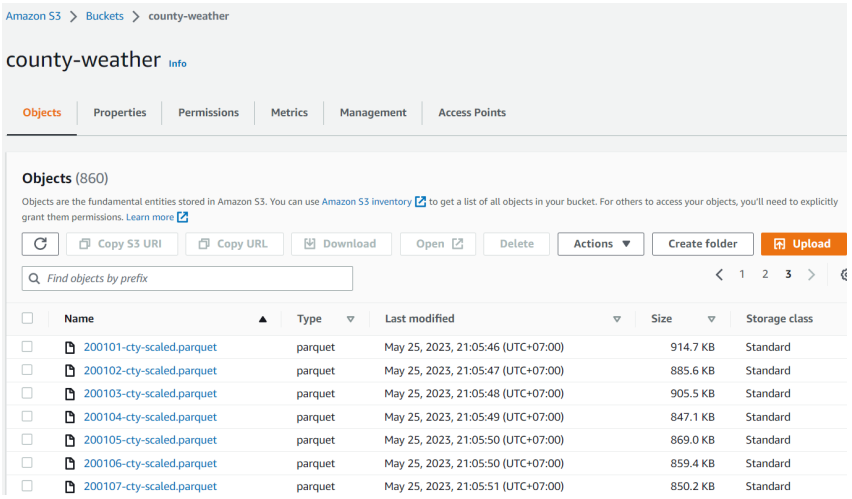
curl -o snow-lab.csv “*https://snow-labb.s3.us-east-1.com/Snow_Lab-snow_*
coverage.csv?....”

Then, we install 2 Python libraries necessary to convert them into PARQUET format: pyarrow and fastparquet. The files will be uploaded to s3 bucket.



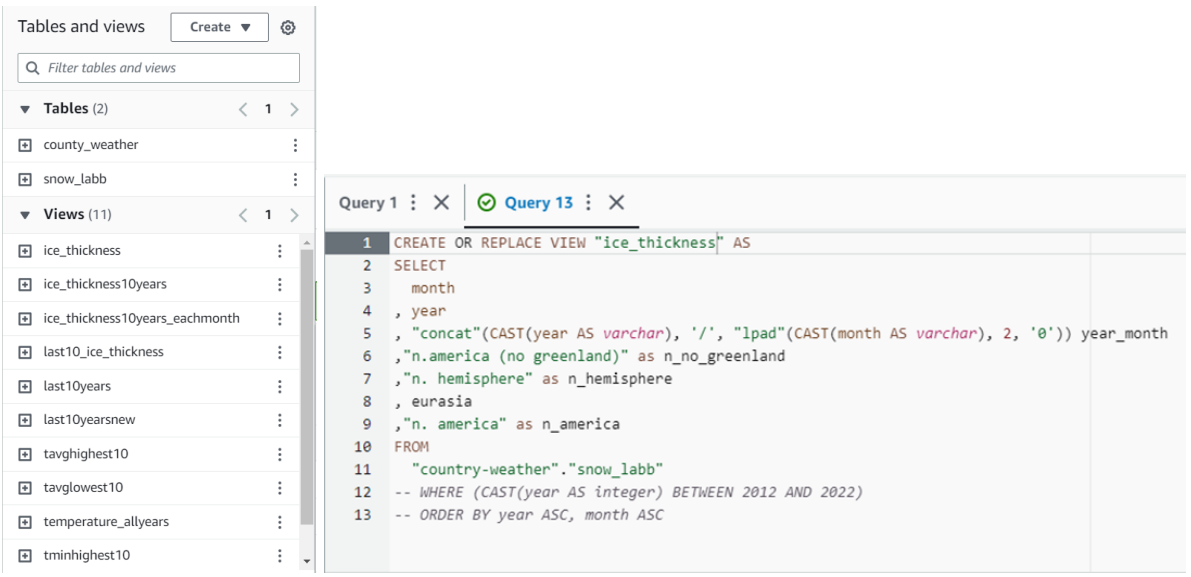
2. Storage

We choose Amazon Simple Storage Service (S3) for storage. We create 2 buckets: county-weather and snow-labb. The following picture shows the files inside the county-weather bucket.



3. Processing and Analysis

We create an AWS Glue database and table by using the Athena query editor, which connects with the datasets in the S3 buckets. With Athena, we can query the data from all files at once. As shown in the picture, there are 2 tables. We create views from queries for data visualization in the end. These SQL scripts are stored in our Github repository.



4. Visualization

Visualization represents data or information in a visual format such as charts, graphs, maps, or infographics. Its primary purpose is to communicate complex data in a clear and intuitive way, enabling users to discern patterns, trends, and relationships within the data.In the context of report creation, AWS's service, Quicksight, is used for visualization. Quicksight is chosen for its capabilities to create compelling visualizations and to deliver operational insights based on datasets from different sources. One such source is Athena, a serverless, interactive query service that makes it easy to analyze data directly in S3 using standard SQL.

In the process of data visualization, it's important to select meaningful metrics to convey the story within the data. For this scenario, we have opted to use the Average values of the data set (e.g., median, maximum, minimum values). This choice is particularly useful when presenting data with vast seasonal variations like temperature across different years. Given that a year comprises multiple data points and each month of the year typically demonstrates a certain seasonality, utilizing the Average value allows us to present an understandable and concise trend.

By utilizing AWS's Quicksight, we can develop more effective, dynamic, and interactive visualizations, which ultimately improves our understanding of the underlying data and aids in data-driven decision making.

Part VI

Result, Conclusion and Discussion

In this study, we examined the proof of concept for using Amazon Web Services (AWS) to utilize big data analytics to tackle global warming and climate change problems. We used a variety of AWS services, including Cloud9, S3 bucket, AWS Glue, Amazon Athena, and Amazon Quicksight, to successfully achieve our objectives. The results of our study showed that AWS can be used to effectively tackle global warming and climate change problems. By using a variety of AWS services, we were able to collect, store, analyze, and visualize large amounts of data to gain a better understanding of the causes and effects of climate change. This information can then be used to develop solutions to mitigate the effects of climate change.

The impact of the U.S. land temperature on snow coverage areas are analyzed by two visualizations, the monthly average U.S. land temperature in each year **(Figure 2)** and the monthly average U.S. area of snow coverage in each year **(Figure 3)**.

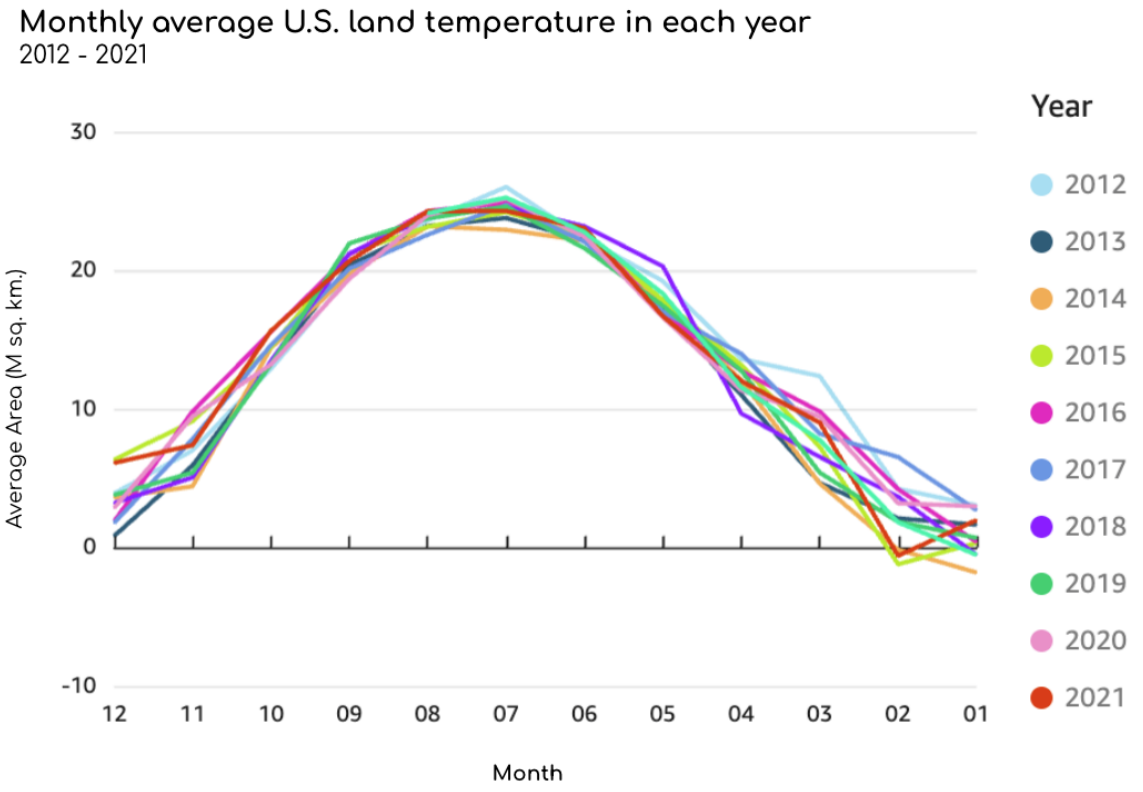


Figure 2. The monthly average U.S. land temperature in each year 2012-2021

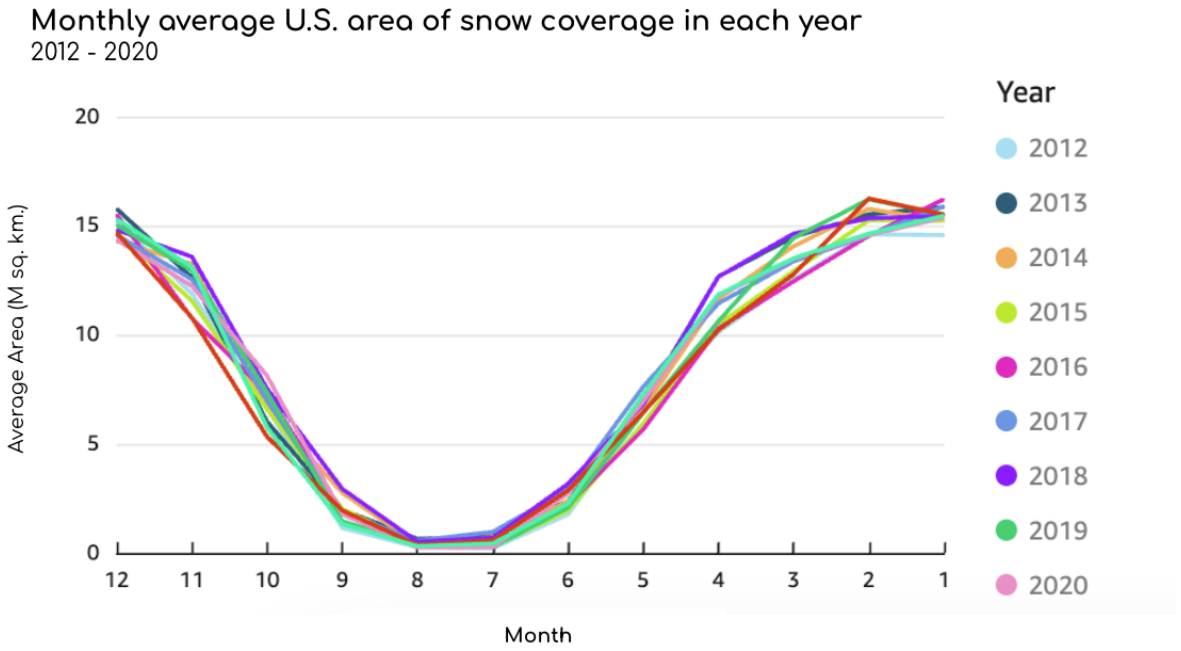


Figure 3. The monthly average U.S. area of snow coverage in each year 2012-2020

The results show that there is a strong negative correlation between temperature and snow coverage areas. As temperature increases, snow coverage areas decrease. As a result of the increasing average global temperature, **Figure 4** shows the trends of decreasing snow coverage in North America.

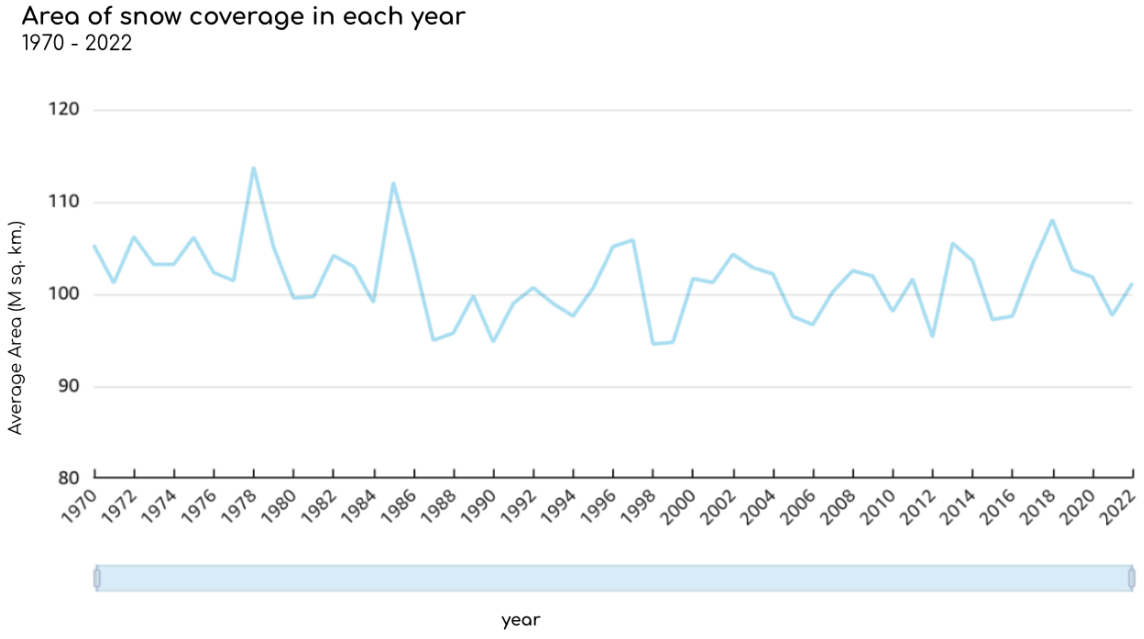


Figure 4. The trends of decreasing snow coverage in North America across 1970-2022

Part VII
Reflection & Suggestion for Further Studies

From this group project work, there are many things that we can learn about using AWS service to deal with the big data project about global warming. Our main key takeaways are:

- **The importance of data collection:** In order to understand the impact of climate change, we need to collect data from a variety of sources. This data can include research, news, weather datasets, and case studies of real-world impact of climate change datasets.
- **The importance of data pipeline designing:** Data pipeline designing is the process of creating a system for collecting, storing, and processing data. A well-designed data pipeline can ensure that data is collected and processed in a timely and efficient manner, and that it is accessible to users when they need it. Data pipeline designing is an essential part of any data-driven organization, and it is especially important for organizations that are working to understand and address climate change.
- **The importance of data analysis and visualization:** To identify trends and patterns. The analysis can help us to understand how climate change is affecting the Earth's activities. Once we have analyzed the data, we need to visualize it in a way that is easy to understand. This visualization can help us to communicate the findings of our research to others.

For further studies, the economical aspect such as tourism revenue and tourist count can be used to analyze more in-depth the effect of global warming and the tourism industry in the U.S. and all regions of the world. For example, a study could be conducted to compare the tourism revenue of a ski resort before and after a period of global warming. The study could also look at how the number of winter tourists has changed over time. This information could be used to better understand the economic impact of global warming and to develop strategies to mitigate its effects.

Part VIII
Project Resource

Git-Hub Repository Link

Link : https://github.com/golf-ratch/climateChangeCS653.git

Presentation slide Link

Link : Presentation slide

Reference

[1] Abbass, K., Qasim, M.Z., Song, H. et al. A review of the global climate change impacts, adaptation, and sustainable mitigation measures. Environ Sci Pollut Res 29, 42539–42559 (2022). https://doi.org/10.1007/s11356-022-19718-6

[2] Rutgers University. Area of Snow Extent, https://climate.rutgers.edu/snowcover/table_area.php. Retrieved May, 2023

[3] I.Gerretsen. How climate change threatens to close ski resorts, https://www.bbc.com/future/article/20230124-how-climate-change-threatens-to-close-ski-resorts Retrieved May, 2023

[4] NOAA U.S. Climate Gridded Dataset (NClimGrid) was accessed on May, 2023 from https://registry.opendata.aws/noaa-nclimgrid.