Climate Change

Claire Lee, Samyu Krishnasamy, Bianca Linares, Semin Ahn

Data Selection - Climate Change: Earth Surface Temperature Data

Global Land Temperatures By Major City

- dt (date)
- Average Temperature
- Average Temperature Uncertainty
- City
- Country
- Latitude
- Longitude

Global Land Temperatures By State

- dt (date)
- Average Temperature
- Average Temperature Uncertainty
- State
- Country

Goal: Analyze long-term climate trends to uncover regional variations in surface temperatures across major cities and states, focusing on:

- Identifying global warming patterns by observing changes in average temperatures over time.
- Comparing temperature trends between urban areas (major cities) and broader regions (states) to understand the impact of urbanization and industrialization.

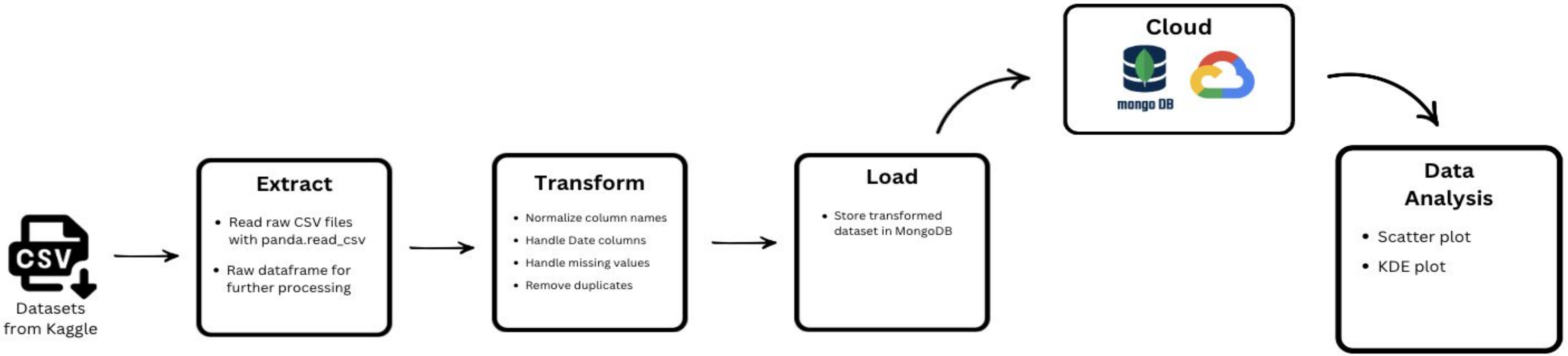
Difficulties

- Finding datasets that were both relevant to the assignment and had enough data.
- Another difficulty was finding a dataset that was made by a credible source

Provenance: Kaggle/Berkeley Earth Surface Temperature Study

https://github.com/claireylee/DataScienceFinalProject

ETL Pipeline



Datasets from Kaggle

Extract

- Read raw CSV files with panda.read_csv
- Raw dataframe for further processing

Transform

- Normalize column names
- Handle Date columns
- Handle missing values
- Remove duplicates

Load

- Store transformed dataset in MongoDB

Cloud

Data Analysis

- Scatter plot
- KDE plot

Cloud Storage

Project Creation:

- Created a Google Cloud project to manage resources and permissions
- Enabled necessary APIs

BigQuery Dataset Setup:

- Navigated to BigQuery Console in the Google Cloud
- Created new datasets to organize and store transformed data

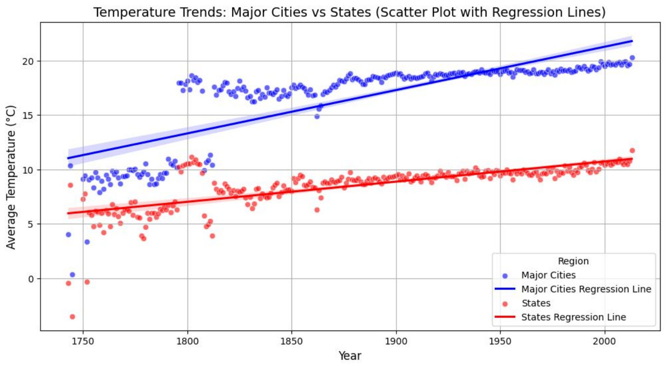
Data Upload:

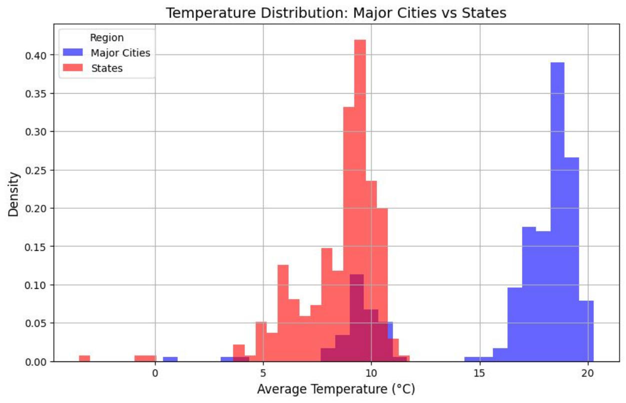
- Uploaded transformed datasets directly to BigQuery tables
- Defined table schemas to match the structure of the transformed data

Data Accessibility:

- Ensured the data is securely stored and accessible for analysis

Analysis


Temperature Trends: Major Cities vs States (Scatter Plot with Regression Lines)


Temperature Distribution: Major Cities vs States

```
          =======Major City=======

      averagetemperature  averagetemperatureuncertainty

count          228175                  228175

mean        18.125969               0.969343

std        10.024800                0.979644

min       -26.772000                0.040000

25%         12.710000                0.340000

50%         20.428000                0.592000

75%         25.918000                1.320000

max         38.283000               14.037000
```

```
          =======State=======

      averagetemperature  averagetemperatureuncertainty

count     620027                   620027.000000

mean        8.993111                    1.287647

std       13.772150                    1.360392

min       -45.389000                    0.036000

25%        -0.693000                    0.316000

50%        11.199000                    0.656000

75%        19.899000                    1.850000

max        36.339000                   12.646000
```

Challenges/Insights

Challenge 1: Managing Large Datasets

- Problem: Extracting and loading large datasets caused memory spikes and performance delays, especially with tools like pd.read_csv().
- Solution: Implemented chunked reading with Python’s pandas to process data in smaller, manageable portions. Used bulk_write() in MongoDB to batch operations, improving insertion speed and efficiency.

Challenge 2: Cloud Integration Issues

- Problem: Establishing and maintaining a connection between Google Cloud and Google Colab was initially confusing, requiring proper authorization and active connections.
- Solution: Generated and managed credentials to ensure seamless integration. Troubleshot workflows to maintain connectivity, improving the pipeline's reliability.

Challenge 3: Duplicate Data Handling

- Problem: Inserting new data into the database often resulted in duplicate records, disrupting consistency.
- Solution: Employed bulk operations with upsert to ensure existing records were updated and new records inserted without duplication. Split data into smaller batches, reducing processing time and improving overall accuracy.

Technical Lessons:

- Scalability: Leveraged chunked processing and distributed systems to handle large datasets effectively.
- Cloud Expertise: Developed skills in integrating Google Cloud with analytical tools like Google Colab for seamless workflows.

Analytical Lessons:

- Visualization: Improved ability to identify and communicate trends and outliers through iterative experimentation.
- Team Coordination: Learned the importance of structured workflows and clear task delegation for project success.