Deliverable – 3
Climate Change: Earth Surface Temperature Data

Team members

Mitra Buggaveeti (801310349)
Deeksha Reddy Ganta (801311836)
Sai Kiran Reddy Bokka
Lahari Prathapagiri (801318004)
Lakshmi Prasanna Adeboyena

Communication plan to include project artifact repository:

Communication channels: Discord for daily team conversation and syncing. We will be using email for formal ways of communication.
Communication frequency: 20 mins long standup call on zoom for every alternate day to check the progress of the project
Work division : Tasks are divided among team members efficiently any work overload can be brought to team notice and discussed to resolve them.
Meeting agendas: Update on tasks assigned to each individual, assigning new tasks and planning future developments, discussing issues faced and collectively working to fix them.

Project repository can be accessed on Github using the following Link given below:
Github link to access the repository:

https://github.com/mitrabuggaveeti/BDA_Project_13/blob/deliverable3/README.md

Data Set Selection :

We have chosen climate change information from Kaggle that clarifies how the planet's surface temperature has changed throughout time. Between the beginning of the 1990s to the present, the temperature has altered dramatically. We will make use of the Berkeley Earth data collection, which has 1.6 billion properly formatted temperature reports from 16 pre-existing archives. The average land temperature and data combining the average land and ocean temperatures can both be found in this dataset.

The Dataset we selected for our modeling and analysis is

https://www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-temperature-data

Business Problem:

The chosen dataset can be used to address various business problems and opportunities such as:

1. Climate change research: The dataset can be used to spot long-term temperature trends and look into how climate change is affecting various parts of the world. This could be helpful for organizations working to mitigate and adapt to climate change, as well as for legislators and advocacy groups.
2. Sustainable development: Businesses, especially those in sectors with high greenhouse gas emissions, might utilize this dataset to inform their sustainable development strategy. Companies can find opportunities for innovation and investment in sustainable technology and practices by studying how temperature is changing in various places and how it may affect business operations.
3. Risk assessment: The dataset can be utilized for disaster preparation and risk assessment as well. Businesses can identify areas most vulnerable to extreme weather events, such as heatwaves, droughts, or flooding, by examining temperature patterns, and create backup plans to reduce these risks.

Input Data Set:

We have selected our dataset from Kaggle. Earth Surface Temperature Data set available on Kaggle is a comprehensive collection of historical temperature records from around the world, spanning over 250 years. The dataset contains monthly temperature measurements from January 1750 to December 2015, covering over 200 countries and major cities worldwide.

Here's an explanation of each attribute in the dataset:

Date: Provides the date

Average Temperature: Gives the average temperature for each city

City: Gives the city name

Country: Gives the country name

Latitude: A coordinate that indicates a point's north-south location of the city

Longitude: A coordinate that indicates a point's east-west location on the city.

Research Objectives:

The agenda of our team is to identify the strategies that could help us define the solutions to the below questions using suitable visualizations as part of our research work. By analysis the chosen dataset we would further investigate the following questions:

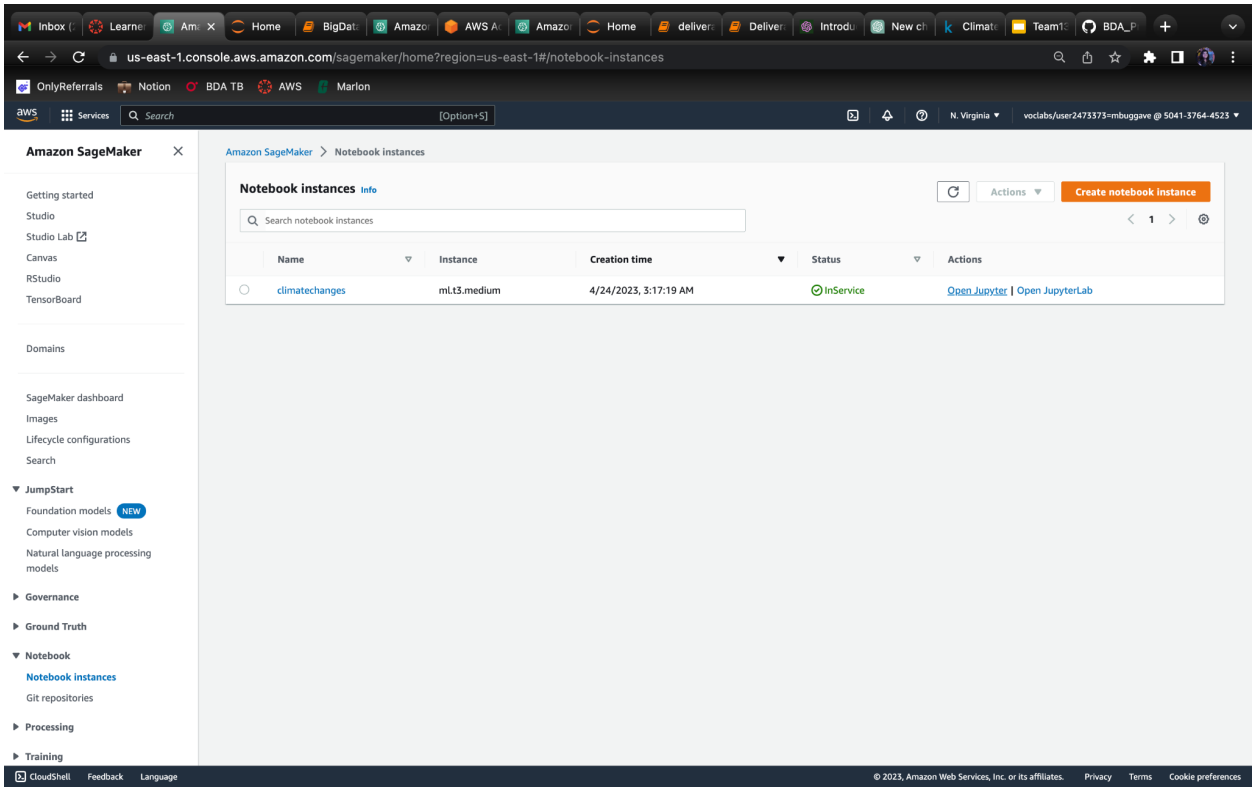
- In correlation to changes in temperature, are there population changes in animals or plants?
- How do temperature variations over different periods, including daily or seasonal variations, relate to long-term changes in average temperature?
- Between changes in temperature and differences in people's health, are there any relationships?
- How can we use machine learning and other advanced analytics techniques to gain insights from the large and complex datasets related to climate change and global temperature anomalies?

7. Analytics and Machine Learning

- This is a sizable dataset that includes historical global temperatures. This data set is frequently used in machine learning and analytics applications for research on climate change, weather forecasting, and several other relevant topics.
- To learn more about the properties of the data, such as patterns, trends, and anomalies in temperature data, exploratory data analysis, or EDA, can be utilized.
- Data analysis can be used to find links and relationships between temperature and other factors like as time, place, and dioxide emissions.
- The use of data visualization tools can be used to produce interactive maps and charts that display temperature patterns over time and in various geographic locations.
- This dataset may be utilized to teach algorithms that use machine learning and develop models of prediction that predict upcoming variations in temperature based on past information.
- Future temperature can be forecast via regression models, particularly the use of linear regression and time-series data models for forecasting.
- In addition to classifying and identifying locations that are most vulnerable to climate change, algorithms like clustering and classification can also be used to group comparable geographic areas based on temperature fluctuations.

Implementations:

1. First we create a jupyter notebook and added the dataset to the S3 bucket named “climatechanges”.



2. To connect the dataset with Jupyter Notebook, we used the following code to read the dataset:

```
In [2]: s3_csv_path = f's3://group13deliverable2/climate/archive/GlobalLandTemperaturesByCity.csv'
cities = pd.read_csv(s3_csv_path)
cities
```

Out[2]:

	dt	AverageTemperature	AverageTemperatureUncertainty	City	Country	Latitude	Longitude
0	1743-11-01	6.068	1.737	Århus	Denmark	57.05N	10.33E
1	1743-12-01	NaN	NaN	Århus	Denmark	57.05N	10.33E
2	1744-01-01	NaN	NaN	Århus	Denmark	57.05N	10.33E
3	1744-02-01	NaN	NaN	Århus	Denmark	57.05N	10.33E
4	1744-03-01	NaN	NaN	Århus	Denmark	57.05N	10.33E
...	...	...	...	...	...	...	...
8599207	2013-05-01	11.464	0.236	Zwolle	Netherlands	52.24N	5.26E
8599208	2013-06-01	15.043	0.261	Zwolle	Netherlands	52.24N	5.26E
8599209	2013-07-01	18.775	0.193	Zwolle	Netherlands	52.24N	5.26E
8599210	2013-08-01	18.025	0.298	Zwolle	Netherlands	52.24N	5.26E
8599211	2013-09-01	NaN	NaN	Zwolle	Netherlands	52.24N	5.26E

8599212 rows × 7 columns

Data Cleaning:

Removing or imputing missing values: The dataset may contain missing values, and it is important to identify and deal with them.

Removing duplicates: The dataset contains duplicate records like repeated temperature values for a single year. So we have removed these duplicate values.

Converting data types: The dataset can include data that needs to be converted to the right format or data types that are in the incorrect format. The data collection was converted to Excel format before being displayed.

Renaming columns: The dataset may contain columns with unclear or inconsistent names, and it is best practice to rename columns to improve clarity and consistency.

Normalizing or scaling data: The dataset might include variables with varying scales, and it might be essential to scale or normalize the data to increase the analysis's precision.

```
In [9]: cities.isnull().sum()

Out[9]: dt                             0
        AverageTemperature             0
        AverageTemperatureUncertainty  0
        City                           0
        Country                        0
        Latitude                       0
        Longitude                      0
        dtype: int64

In [10]: duplicate = cities[cities.duplicated()]
         print("Duplicate rows:", len(duplicate))
         print(duplicate)

         Duplicate rows: 0
         Empty DataFrame
         Columns: [dt, AverageTemperature, AverageTemperatureUncertainty, City, Country, Latitude, Longitude]
         Index: []
```

3. We have used linear regression for modeling the data

Train and Test Split of dataset into 80:20

```
In [22]: X_train,X_test,y_train,y_test = train_test_split(X,y,train_size=0.2,random_state=42)
         pred_model= LinearRegression()
```

Model Fitting

In [23]:
```
pred_model.fit(X_train,y_train)
model = pred_model.predict(X_test)
cities['pred'] = pred_model
print("The prediction values of the temperature are" ,model)
```

```
The prediction values of the temperature are [18.12377366 17.34975148 17.71639357 18.01174414 17.4108585  17.76731608
 18.02192864 17.98119063 17.78768508 17.421043   18.30709471 17.67565556
 18.15432717 18.44967774 17.49233452 17.30901347 18.04229764 17.57381053
 18.13395817 17.8487921  18.35801722 18.23580319 17.4312275  17.62473305
 17.99137513 18.43949324 18.40893973 17.53307253 17.75713158 18.39875523
 18.10340466 17.46178101 17.61454854 18.36820172 17.40067399 17.64510205
 17.94045262 17.97100613 17.59417954 17.73676257 17.35993598 17.8691611
 17.96082162 17.70620907 17.66547106 17.47196551 18.33764821 17.65528655
 17.87934561 17.38030499 17.74694708 18.00155963 18.17469617 17.58399504
 17.50251902 18.22561869 18.2765412  17.56362603 17.39048949 18.16451167
 17.80805409 17.441412   18.11358916 17.55344153 17.33956698 17.48215001
 17.69602456 18.25617219 17.37012049 18.05248215 18.2969102  18.21543418
 17.8589766  17.81823859 18.08303565 17.77750058 18.09322016 17.93026812
 18.42930873 18.20524968 17.72657807 17.89971461 17.79786959 18.2867257
 17.88953011 18.07285115 17.63491755 18.41912423 17.90989911 17.95063712
 18.2663567  17.68584006]
```

Mean squared error (MSE)

In [30]:
```
y_pred = pred_model.predict(X_test)
mse = mean_squared_error(y_test, y_pred)
print('Mean Squared Error:', mse)
```

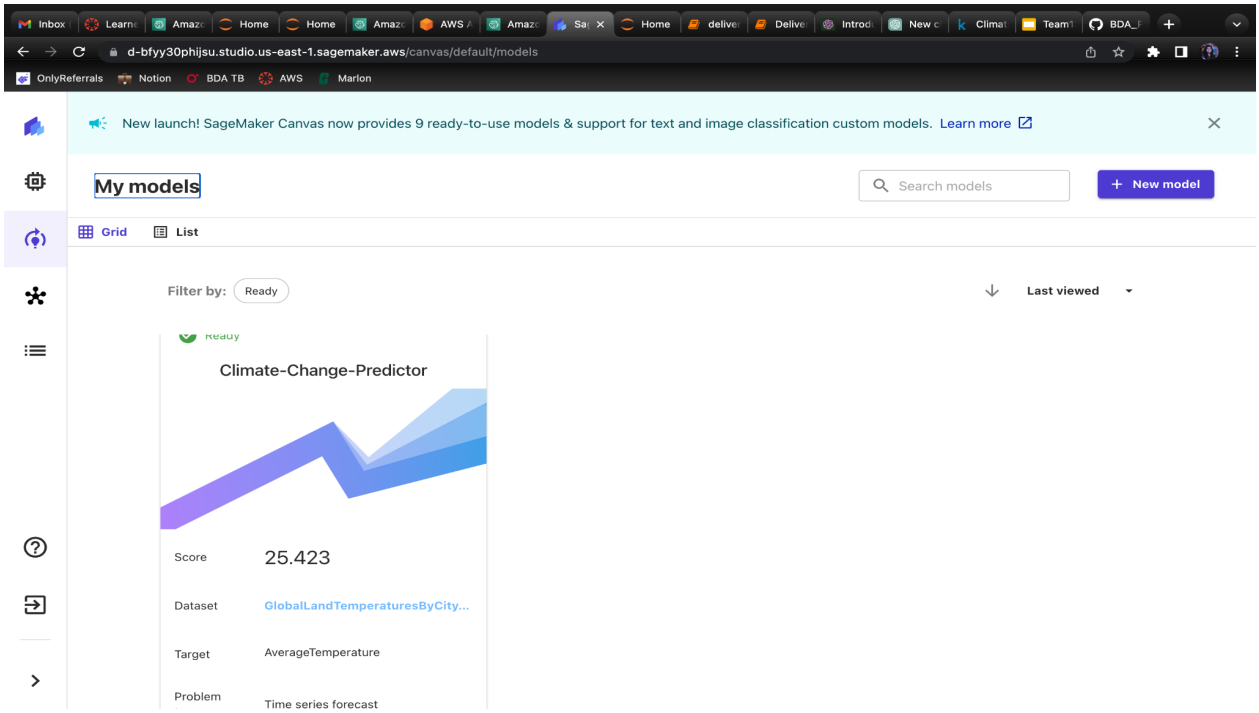
```
Mean Squared Error: 0.03934226585629664
```

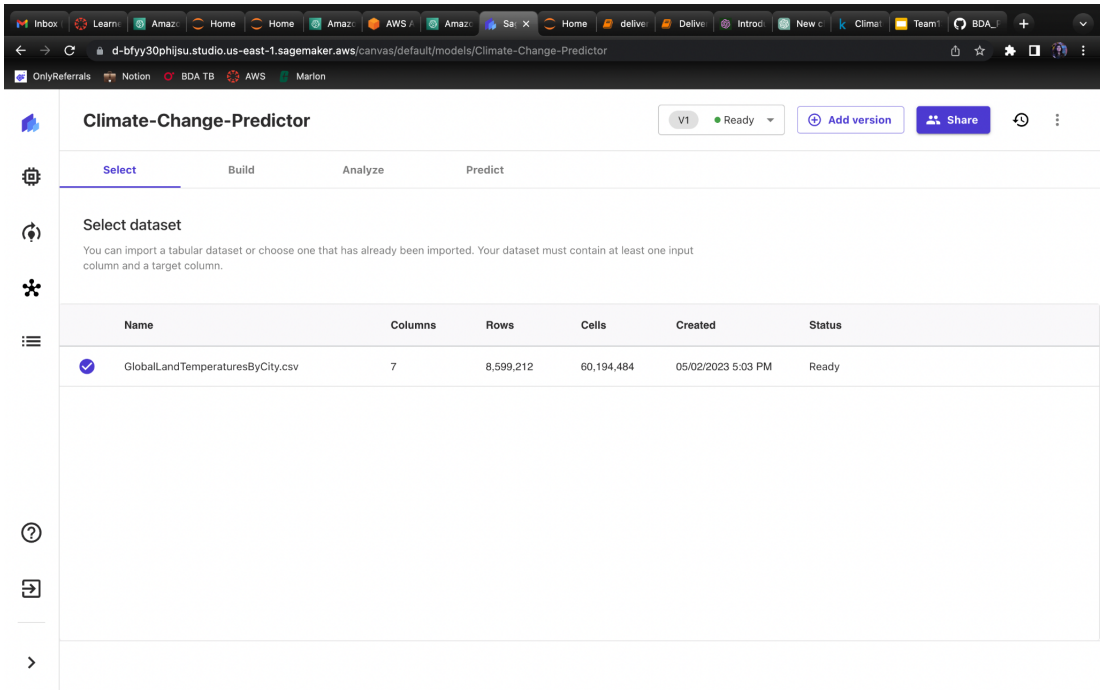
R-squared and adjusted R-squared values for the linear regression model.

In [32]:
```
from sklearn.metrics import r2_score

y_pred = pred_model.predict(X_test)
r_squared = r2_score(y_test, y_pred)
print('R-squared:', r_squared)

n = len(X_test) # number of samples
p = 1 # number of independent variables
adj_r_squared = 1 - (1 - r_squared) * ((n - 1) / (n - p - 1))
print('Adjusted R-squared:', adj_r_squared)
```

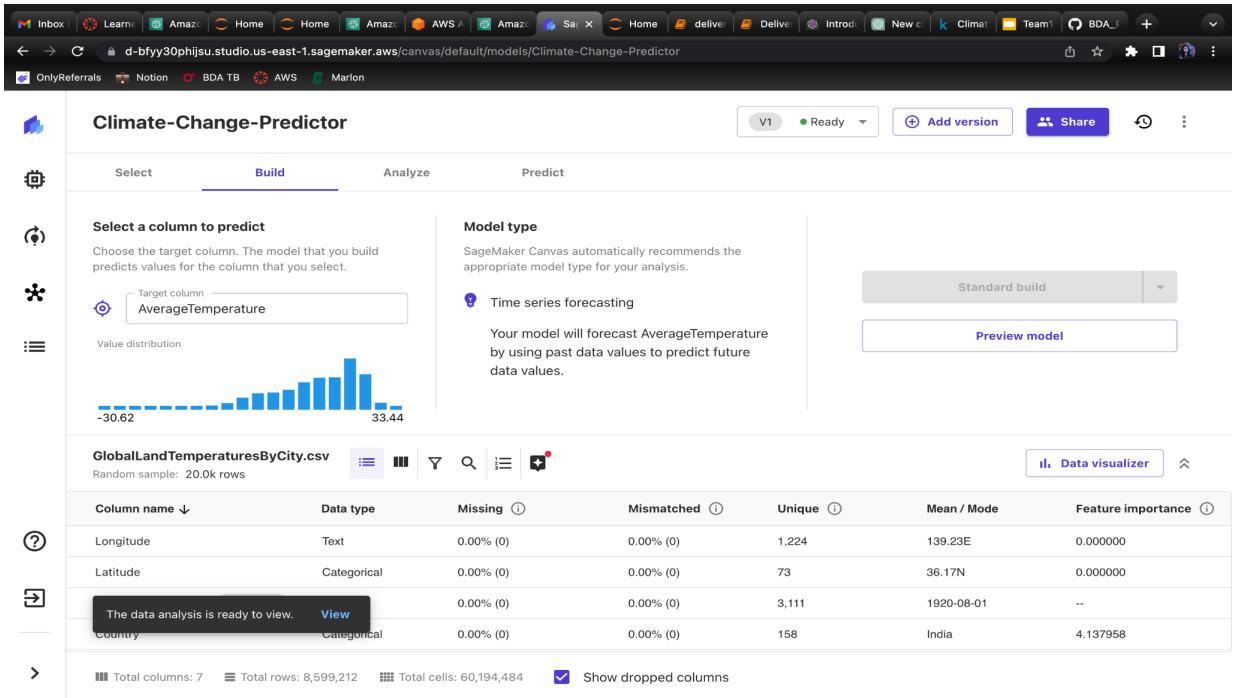
```
R-squared: 0.7109131654750815
Adjusted R-squared: 0.7077010895359158
```

4. **We have used Amazon Sage Maker canvas to implement time series analysis**
   Amazon SageMaker Canvas is a component of Amazon SageMaker that allows you to build, visualize, and monitor machine learning workflows. It provides a visual way to design and build machine learning models and data processing pipelines. SageMaker Canvas also provides real-time monitoring of your workflows, so you can see the progress of your training jobs, track model performance, and get alerts if any issues arise. SageMaker Canvas also provides tools for model evaluation, such as confusion matrices, ROC curves, precision-recall curves, and performance metrics. With Amazon SageMaker Canvas, one of its features, you can create, display, and manage machine learning processes. It offers a visual way to create pipelines for data processing and prediction models. SageMaker Canvas also offers real-time workflow monitoring so you can keep an eye on model performance, track the status of your training jobs, and receive warnings when problems develop. Additionally available through SageMaker Canvas are tools for evaluating models, including confusion matrices, ROC curves, precision-recall curves, and performance metrics.



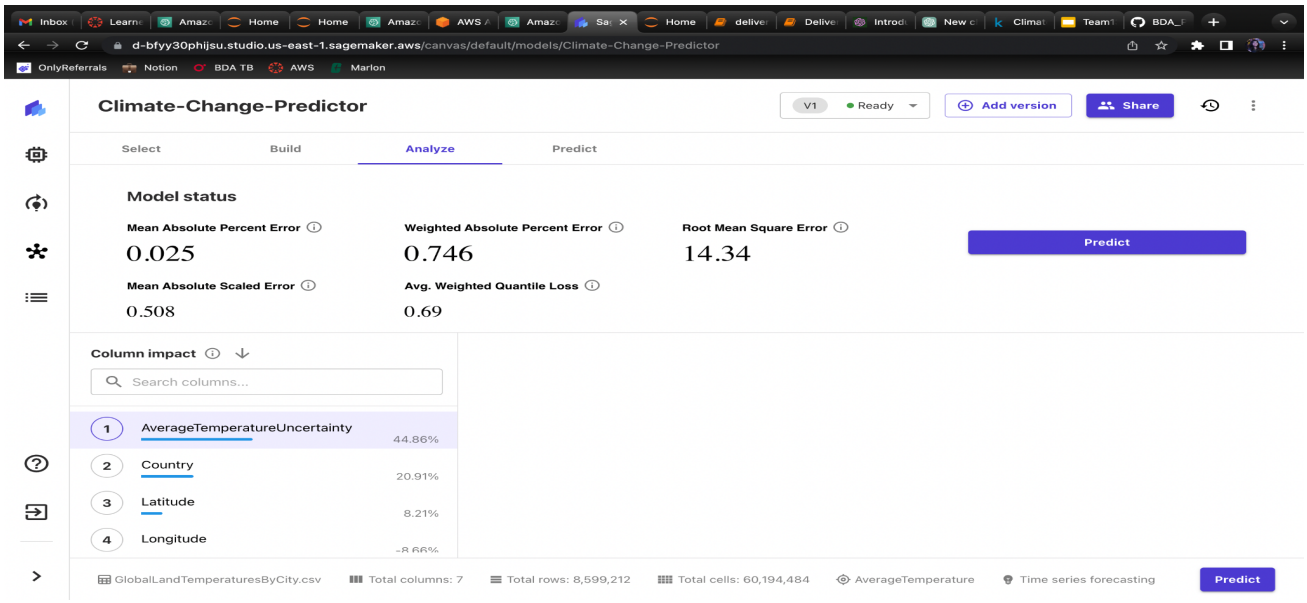


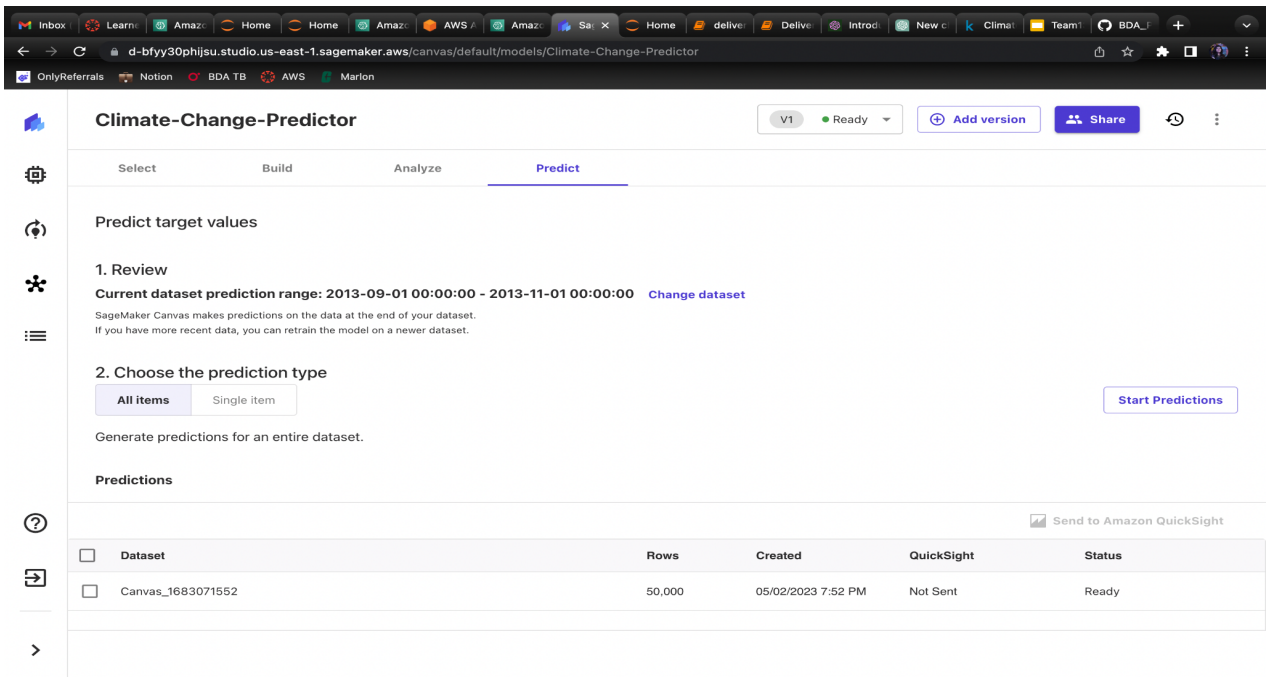
For our model, we used Time series forecasting to predict the AverageTemperature. Once the model is created, various columns are created such as Latitude, Longitude and Country. When you click on each of these, you can see their correspondent relationship values.



Evaluation Metrics:

The evaluation metrics for the model we used are the RMSE values and the percentage error. It provides a measure of the average magnitude of the errors between the predicted and actual values, with a lower value indicating better accuracy. Our model has an RMSE value of 14.34.

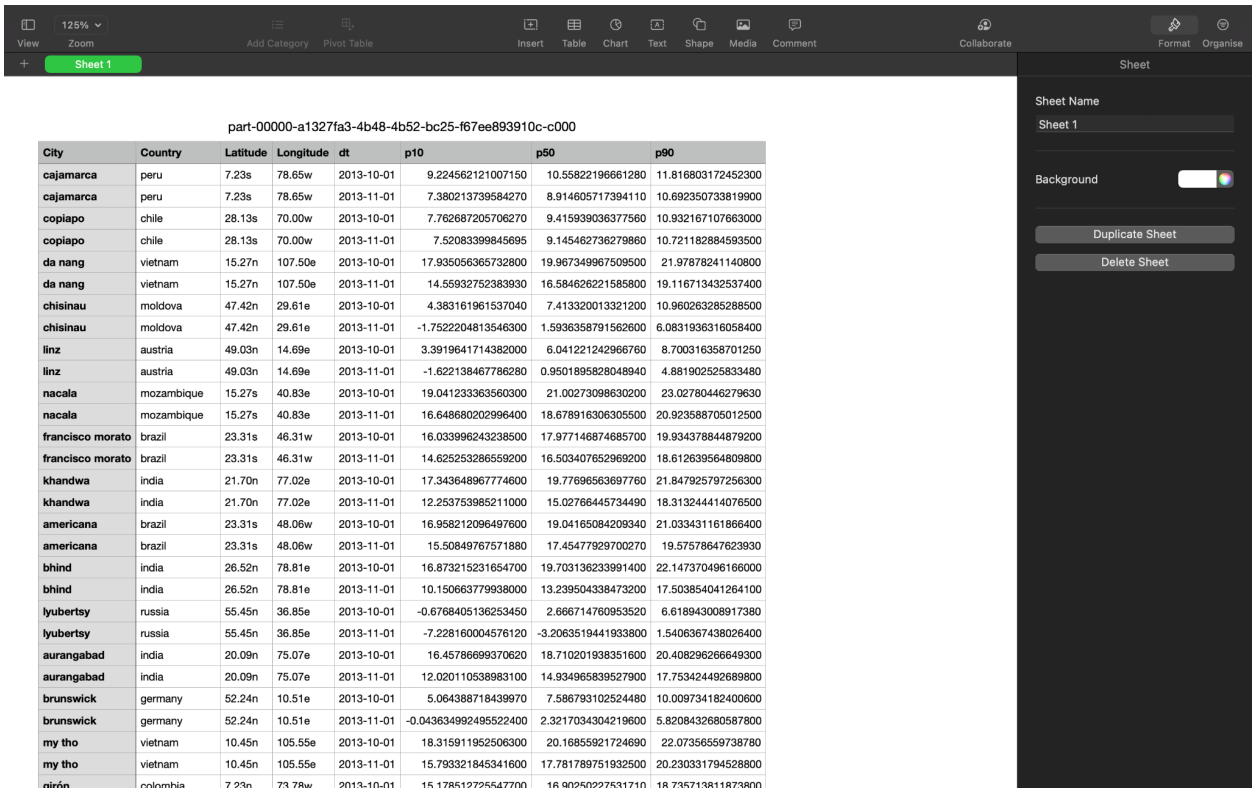




Result:

Once the data model is built, a CSV file is created and three average temperature values are generated in three columns for the months October and November on the first day. The temperature can lie between any of those three values.

Generated sample temperature predictions are:



part-00000-a1327fa3-4b48-4b52-bc25-f67ee893910c-c000

City	Country	Latitude	Longitude	dt	p10	p50	p90
cajamarca	peru	7.23s	78.65w	2013-10-01	9.224562121007150	10.55822196661280	11.816803172452300
cajamarca	peru	7.23s	78.65w	2013-11-01	7.380213739584270	8.914605717394110	10.692350733819900
copiapo	chile	28.13s	70.00w	2013-10-01	7.762687205706270	9.415939036377560	10.932167107663000
copiapo	chile	28.13s	70.00w	2013-11-01	7.52083399845695	9.145462736279860	10.721182884593500
da nang	vietnam	15.27n	107.50e	2013-10-01	17.935056365732800	19.967349967509500	21.97878241140800
da nang	vietnam	15.27n	107.50e	2013-11-01	14.55932752383930	16.584626221585800	19.116713432537400
chisinau	moldova	47.42n	29.61e	2013-10-01	4.383161961537040	7.413320013321200	10.960263285288500
chisinau	moldova	47.42n	29.61e	2013-11-01	-1.7522204813546300	1.5936358791562600	6.083193636058400
linz	austria	49.03n	14.69e	2013-10-01	3.3919641714382000	6.041221242966760	8.700316358701250
linz	austria	49.03n	14.69e	2013-11-01	-1.622138467786280	0.9501895828048940	4.881902525833480
nacala	mozambique	15.27s	40.83e	2013-10-01	19.041233363560300	21.00273098630200	23.02780446279630
nacala	mozambique	15.27s	40.83e	2013-11-01	16.648680202996400	18.678916306305500	20.923588705012500
francisco morato	brazil	23.31s	46.31w	2013-10-01	16.033996243238500	17.977146874685700	19.934378844879200
francisco morato	brazil	23.31s	46.31w	2013-11-01	14.625253286559200	16.503407652969200	18.612639564809800
khandwa	india	21.70n	77.02e	2013-10-01	17.343648967774600	19.77696563697760	21.847925797256300
khandwa	india	21.70n	77.02e	2013-11-01	12.253753985211000	15.02766445734490	18.3132444140765000
americana	brazil	23.31s	48.06w	2013-10-01	16.958212096497600	19.04165084209340	21.033431161866400
americana	brazil	23.31s	48.06w	2013-11-01	15.50849767571880	17.45477929700270	19.57578647623930
bhind	india	26.52n	78.81e	2013-10-01	16.873215231654700	19.703136233991400	22.147370496166000
bhind	india	26.52n	78.81e	2013-11-01	10.150663779938000	13.239504338473200	17.503854041264100
lyubertsy	russia	55.45n	36.85e	2013-10-01	-0.6768405136253450	2.666714760953520	6.618943008917380
lyubertsy	russia	55.45n	36.85e	2013-11-01	-7.228160004576120	-3.2063519441933800	1.5406367438026400
aurangabad	india	20.09n	75.07e	2013-10-01	16.45786699370620	18.710201938351600	20.408296266649300
aurangabad	india	20.09n	75.07e	2013-11-01	12.020110538983100	14.934965839527900	17.753424492689800
brunswick	germany	52.24n	10.51e	2013-10-01	5.064388718439970	7.586793102524480	10.009734182400600
brunswick	germany	52.24n	10.51e	2013-11-01	-0.043634992495522400	2.3217034304219600	5.8208432680587800
my tho	vietnam	10.45n	105.55e	2013-10-01	18.315911952506300	20.16855921724690	22.07356559738780
my tho	vietnam	10.45n	105.55e	2013-11-01	15.793321845341600	17.781789751932500	20.230331794528800
girón	colombia	7.23n	73.78w	2013-10-01	15.178512725547700	16.902502275317100	18.735713811873800

Future

In the dataset, the values are displayed till 2013 (September). So we have enhanced it so that we can predict the temperature values for the next few months (October, and November). To more fully comprehend the causes and consequences of climate change, build and validate climate models utilizing the dataset. a variety of historical data and the current climate, this may involve using machine learning algorithms for predicting future variations in temperature. We also plan to understand the causes and effects of climate change better, construct and validate climate projections based on the dataset. using past data and climate variables, this could entail using machine learning algorithms to predict upcoming variations in temperature.

Questions :

1. What was unique about the data?

The dataset contains temperature data from thousands of cities and regions worldwide, dating back to 1750. This long-term global coverage is valuable for researchers studying climate patterns and trends over time. The temperature data in the dataset has been carefully cleaned and quality-checked to ensure accuracy and consistency. This helps to reduce errors and biases that can arise from data collection and processing.

In addition to temperature readings, the dataset includes other variables such as latitude, longitude, and country/region information. This allows researchers to analyze the data in more detail and consider other factors that may impact temperature patterns.

2. Did you have to deal with imbalance? What were the problems you faced? How did you solve them?

Spatial Imbalance: The spatial distribution of the dataset is not symmetrical, with some locations having more data than others. This may result in skewed temperature patterns, especially in areas where data is few. To resolve this, we can produce more thorough temperature records by using spatial interpolation techniques to fill in the gaps left by missing data.

Temporal Imbalance: Additionally, there is a temporal imbalance in the collection, with more current data having better spatial and temporal resolution than older data. This can make it difficult to compare long-term temperature trends and gauge how much the climate is changing. We can address this by using statistical techniques that take into account variations in data coverage and quality over time.

3. What data cleaning did you do?

The following steps might be taken to clean and preprocess the data:

1. Removing or imputing missing values: The dataset may contain missing values, and it is important to identify and deal with them.
2. Removing duplicates: The dataset contains duplicate records like repeated temperature values for a single year. So we have removed these duplicate values.
3. Converting data types: The dataset may contain data in the wrong format or data types that need to be converted to the appropriate format. We have converted the data set into Excel format and then visualized it.
4. Renaming columns: The dataset may contain columns with unclear or inconsistent names, and it is best practice to rename columns to improve clarity and consistency.
5. Normalizing or scaling data: The dataset may contain variables with different scales, and it may be necessary to normalize or scale the data to improve the accuracy of the analysis

4. Outlier treatment?

Data points known as outliers differ greatly from the rest of the data in a dataset. Outliers may appear in the "Global Land and Ocean Temperature Anomalies" dataset as a result of measurement errors, climatic fluctuations naturally occurring, or other causes. Any data analysis or modeling process must include handling outliers since they can distort statistical results and impair the precision of predictive models.
The specific strategy utilized will depend on the objectives and specifications of the analytic or modeling work at hand. Statistical methods, domain knowledge, and machine learning are some of the approaches that are frequently combined in this treatment. The removal of too many or too few outliers can have an impact on the quality and dependability of statistical results and prediction models, hence outlier treatment should be done cautiously.

5. Imputation?

Imputation is the process of utilizing statistical or machine learning techniques to fill in the missing values in a dataset.Missing values in this dataset can happen for a number of reasons, including data collection problems or unrelated variables.The accuracy of prediction models can be impacted by missing values, which makes imputation a crucial step in the preparation of data. It is important to use imputation cautiously since doing so can produce biased findings, which can undermine the reliability of statistical inferences and forecasting models.

6. Did you create any new additional features/variables?

In the dataset, the values are displayed till 2013 (September). So we have enhanced it so that we can predict the temperature values for the next few months (October, and November)

7. What was the process you used for evaluation? What was the best result?

The RMSE values and the percentage error are the evaluation measures for the model we employed. A lower value denotes greater accuracy, and it provides a measure of the average magnitude of the errors between the anticipated and actual values. The RMSE of our model is 14.34.

8. What future work would you like to do?

The first step to improving your climate change model is to collect more data. This can include historical climate data, satellite imagery, or any other relevant data sources.

9. Instructions for individuals that may want to use your work

The dataset is chosen from Kaggle and anyone who wishes to use our work needs to set up the AWS environment with appropriate credentials. Also, it might ask for charges which need to be paid before using AWS services. Also, the user must load the Kaggle data into AWS and then set up the AWS SageMAker to run the python notebook consisting of code. All the other requirements for the procedure are mentioned within the github repository.