Fidelity International - Data Analytics

GitHub URL (insert URL here) Github: https://github.com/RinkeeSharma/UCDPA_rinkee_sharma

Abstract: has taken two databases from Kaggle first is climate change and the second database is USA housing

In Database 1: Climate change is a combination of two databases; Global temperature and Global land and temperature by countries
Where we have calculated average temperature for each country created figures using Plotly library. 'Seaborn' as well as 'matplotlib' library is used to create horizontal bars for each country
The calculated mean land temperature in countries, mean land temperature on the continents and mean land temperature in the world

In database USA housing: we have the following 7 columns and 5000 rows
1:Area Population'
2:Avg. Area Number of Bedrooms'
3:'Avg. Area Number of Rooms'
4:'Avg. Area House Age'
5:'Avg. Area Income'
6:'Price'
7:Address

Introduction

Dataset

Dataset 1: Climate Change
URL: https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data/code

Description: Mapping of average temperatures in the countries

It is noticed that Russia has one of the lowest average temperatures (like a Canada). The lowest temperature in Greenland (it is distinctly visible on the map). The hottest country in Africa, on the equator. It is quite natural.
Sort the countries by the mean temperature and plot the Horizontal bar
Used 'seaborn' as well as 'matplotlib' library to create horizontal bars for each country

I order to do the analysis of global warming i have taken data from 'GlobalTemperature.csv' file which has a monthly Earth's temperature and plot it on the chart.

Mean land temperature in the world and mean land temperature on the continents
From the charts, it has been observed that there is global warming nowadays.
The average temperature of the earth Surface has had the highest value in the last three centuries. The fastest temperature growth occurred in the last 30 years!. This indicates id that humanity will fully switch to ecological sources of energy, which will reduce CO2. if not, we will be in a disaster. This Charts also have confidence intervals, which shows that measurement of temperature has become more accurate in the last few years.

The chart of annual temperature changes in certain continents (we take into consideration one country per continent and mark Greenland as the coldest place on Earth).

We can see that since 1980 there has been a continuous increase in mean annual temperature for the countries, which we take into consideration (particularly strong dynamics can be seen in the cold countries). The interruption of the temperature values on the chart is due to the lack of observations in these years.

Dataset2: https://www.kaggle.com/fatmakursun/supervised-unsupervised-learning-examples/notebook

Machine learning is an application of artificial intelligence (AI) that provides systems with the ability to automatically learn and improve from experience without being explicitly programmed.

Some machine learning methods

Machine learning algorithms are often categorized as supervised or unsupervised.

Supervised machine learning
unsupervised machine learning.
Semi-supervised machine learning

Regression Analysis

Regression analysis is a reliable method of identifying which variables have impact on a topic of interest. In order to understand regression analysis fully, it’s essential to comprehend the following terms:

- **Dependent Variable**: This is the main factor that you’re trying to understand or predict.
- **Independent Variables**: These are the factors that you hypothesize have an impact on your dependent variable.

We will use the USA housing dataset for regression prediction.

The data contains the following columns:

- 'Avg. Area Income': Avg. Income of residents of the city house is located in.
- 'Avg. Area House Age': Avg Age of Houses in the same city
- 'Avg. Area Number of Rooms': Avg Number of Rooms for Houses in the same city
- 'Avg. Area Number of Bedrooms': Avg Number of Bedrooms for Houses in the same city
- 'Area Population': Population of city house is located in
- 'Price': Price that the house sold at
- 'Address': Address for the house

Regression Model

We will need to first split up our data into an X array that contains the features to train on, and a y array with the target variable, in this case, the Price column. We will toss out the Address column because it only has text info that the linear regression model can't use.

X and y arrays

Interpreting the coefficients:

Holding all other features fixed, a 1 unit increase in Avg. Area Income is associated with an *increase of $21.52*.
Holding all other features fixed, a 1 unit increase in Avg. Area House Age is associated with an *increase of $164883.28*.
Holding all other features fixed, a 1 unit increase in Avg. Area Number of Rooms is associated with an *increase of $122368.67*.
Holding all other features fixed, a 1 unit increase in Avg. Area Number of Bedrooms is associated with an *increase of $2233.80*.
Holding all other features fixed, a 1 unit increase in Area Population is associated with an *increase of $15.15*

Regression Evaluation Metrics

Here are three common evaluation metrics for regression problems:
Mean Absolute Error (MAE) is the mean of the absolute value of the errors:
|
Mean Squared Error (MSE) is the mean of the squared errors:
Root Mean Squared Error (RMSE) is the square root of the mean of the squared errors:

Comparing these metrics:

- **MAE** is the easiest to understand, because it's the average error.
- **MSE** is more popular than MAE, because MSE "punishes" larger errors, which tends to be useful in the real world.
- **RMSE** is even more popular than MSE, because RMSE is interpretable in the "y" units.

All of these are **loss functions** because we want to minimize them.

K Means Clustering with Python

K Means Clustering is an unsupervised learning algorithm that tries to cluster data based on their similarity. Unsupervised learning means that there is no outcome to be predicted, and the algorithm just tries to find patterns in the data. In k means clustering, we have the specify the number of clusters we want the data to be grouped into. The algorithm randomly assigns each observation to a cluster, and finds the centroid of each cluster. Then, the algorithm iterates through two steps: Reassign data points to the cluster whose centroid is closest. Calculate new centroid of each cluster. These two steps are repeated till the within cluster variation cannot be reduced any further. The within cluster variation is calculated as the sum of the euclidean distance between the data points and their respective cluster centroids.

Insights (Point out at least 5 insights in bullet points)

1: head USAHousing

```
Data columns (total 7 columns):
Avg. Area Income                5000 non-null float64
Avg. Area House Age             5000 non-null float64
Avg. Area Number of Rooms       5000 non-null float64
Avg. Area Number of Bedrooms    5000 non-null float64
Area Population                 5000 non-null float64
Price                           5000 non-null float64
Address                         5000 non-null object
dtypes: float64(6), object(1)
```

2: Describe USA housing

Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price	
count	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5.000000e+03
mean	68583.108984	5.977222	6.987792	3.981330	36163.516039	1.232073e+06

std	10657.991214	0.991456	1.005833	1.234137	9925.650114	3.531176e+05
min	17796.631190	2.644304	3.236194	2.000000	172.610686	1.593866e+04
25%	61480.562388	5.322283	6.299250	3.140000	29403.928702	9.975771e+05
50%	68804.286404	5.970429	7.002902	4.050000	36199.406689	1.232669e+06
75%	75783.338666	6.650808	7.665871	4.490000	42861.290769	1.471210e+06
max	107701.748378	9.519088	10.759588	6.500000	69621.713378	2.469066e+0

3: Global Land and Ocean-and-Land Temperatures (GlobalTemperatures.csv):

- Date: starts in 1750 for average land temperature and 1850 for max and min land temperatures and global ocean and land temperatures
- LandAverageTemperature: global average land temperature in celsius
- LandAverageTemperatureUncertainty: the 95% confidence interval around the average
- LandMaxTemperature: global average maximum land temperature in celsius
- LandMaxTemperatureUncertainty: the 95% confidence interval around the maximum land temperature
- LandMinTemperature: global average minimum land temperature in celsius
- LandMinTemperatureUncertainty: the 95% confidence interval around the minimum land temperature
- LandAndOceanAverageTemperature: global average land and ocean temperature in celsius
- LandAndOceanAverageTemperatureUncertainty: the 95% confidence interval around the global average land and ocean temperature

Other files include:

- Global Average Land Temperature by Country (GlobalLandTemperaturesByCountry.csv)

4: Mapping of average temperatures in the countries and sort the countries by the mean temperature and plotted Horizontal bar

5: From the charts it can be observed, that there is global warming nowadays. The average temperature of Earth surface has the highest value in the last three centuries. The fastest temperature growth occurred in the last 30 years! This indicates that if humanity will fully switch to ecological sources of energy, that will reduce CO2. If not, we will be in disaster.This charts also have confidence intervals, which shows that measurement of temperature has become more accurate in the last few years

6: We can see that since the 1980 there has been a continuous increase in mean annual temperature for the countries, which we take into consideration (particularly strong dynamics can be seen in the cold countries). The interruption of the temperature values on the chart is due to the lack of observations in these years.

6:Conclusion of dataset 1

During our project, it was found that there has been a global increase trend in temperature, particularly over the last 30 years. This is due to the violent activities of humankind. In more developed countries the increase in temperature began to register much earlier. Over time the accuracy of the observations is increased, which is quite natural. Mankind must reflect and take all necessary remedies to reduce emissions of greenhouse gases in the atmosphere. This work was entirely done using python. We have received practical skills and knowledge in python and in the library for data visualisation through thus project. The basics of data analytics learned in the elective course helped us in better understanding the project and also helped us in writing a successful python code.