FINAL PROJECT (STAT 517)

Yadav, Rohit Kumar (yada6101@vandals.uidaho.edu)
Masters Computer Science, University of Idaho

Climate Analysis and Temperature Pridictions

Abstract

Climate change poses real and present danger to the future of the world. Also at stake is practically every major life form's existence that resides on this planet. Primarily, addressing climate change requires developing awareness in peoples' minds**.** This Machine Learning Project is based on climate datasets. climate statistics and forecast is an important resource that the government has not explored commensurate to its impact.

The aim of this project is to make this process computerized by implementing principles of data mining and analytics by implementing various types of regression models and clustering. More specifically, this project aims at targeting the social issue of climate condition, analysing data based on climate and temperature, amount of rainfall, snow, thunder and similar factors for different regions in New Delhi, India.

Data can be mined and analysed to find various trends and relations, such as – comparison between climate conditions in area; and can try to predict the temperature on the basis of climate condition in different regions using regression. Clustering can be used to find features where the climate conditions are mostly similar.

The end result of the project will be research based reports specifying these trends, studied and analysed from data taken over the past few years in New Delhi region on the dataset of 100990 of rows and 20 columns.My previous preliminary testing was on the agricultural datasets which was not big enough datasets to perform all the models and clustering visualizations so I switched it with similar kind of data but in other region whith huge datasets of 100990 of rows and 20 columns.

FINAL PROJECT (STAT 517)

Yadav, Rohit Kumar (yada6101@vandals.uidaho.edu)
Masters Computer Science, University of Idaho

Background

It is estimated that climate vulenreability will displace 250 million people by 2050 [1]. A more shocking fact is that if everyone in the world lived the way people do in the U.S, it would take five Earths to provide enough resources for everyone [2]. Additionally, thirty seven percent of Americans believe that global warming is a hoax, and 64 percent don't believe that climate change will seriously affect their way of life. [3] The greater problem is how nations choose to tackle this problem of climate change. There have been around 2,950,000 publications on climate change according to Google Scholar. Although most of them produce interesting results, a majority of them fail to attract attention of the general populace through simple media like interactive visualizations. We intend to achieve a part of this function through the present experiment.

Introduction

Climate change poses real and present danger to the future of the world. Also at stake is practically every major life form's existence that resides on this planet. Primarily, addressing climate change requires developing awareness in peoples' minds. We have done cluster analysis on the similar climate conditions and implemented different types of regression models contributing to climate change, and produced several visualizations which establish relations between several socio-economic factors. More generally, we have created visualizations to trigger climate-aware thought-processes and provides some useful insights through basic statistical and machine learning methods.

Data Mining is an emerging research field in weather analysis. In this project, our focus is on the applications of Data Mining techniques in agricultural field. Different Data Mining techniques are in use, such as K-Means, K-Nearest Neighbour (KNN) and Support Vector Machines (SVM) for very recent applications of Data Mining techniques in agricultural field. In this project, consider the problem of predicting yield temperatures. Yield temperatures is a very important weather problem that remains to be solved based on the available data. The problem of yield prediction can be solved by employing Data Mining techniques. This work aims at finding suitable data models that achieve a high accuracy and a high generality in terms of yield prediction capabilities. For this purpose, different types of Data Mining techniques were evaluated on data sets.

The project aims at performing a thorough analysis of Climate Change and temperature datasets made available by the Delhi-government of India found on the Kaggle. We found many relations between factors affecting climate change, and produce thought-provoking visualizations to increase awareness. Also, through the power of learning theory, We analysed the factors contributing to climate change, and predicts future conditions if human actions proceed as they do today. We rendered many interesting visualizations providing vivid representations and run-throughs of data that has been collected over many years by different sources. We produced interactive plots and clusters that show different factors against each other and the climate conditions.

Why This is an Important Problem?

- Nowadays many research on climate change has been limited to papers and mathematical figures where It fails to provide interactive visualizations for people to see so Possibly, there are interesting relationships between factors leading to or affecting climate change. It would be interesting to find out these relations.
- Machine Learning methods can allow us to model the conditions and predict the future.

The Datasets

we have used the climate dataset from the Delhi-government of India found on Kaggle which consists of of 100990 of rows and 20 columns. It consists of Delhi-weather data from 1997 to 2016 december. This data was taken out from wunderground with the help of their easy to use api. It contains various features such as temperature, pressure, humidity, rain, precipitation,etc.

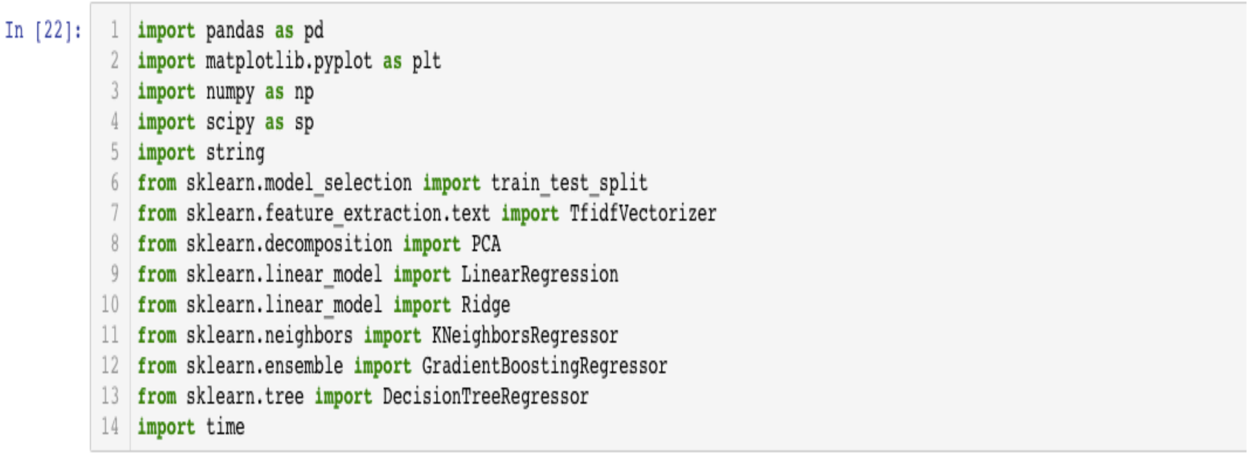
FINAL PROJECT (STAT 517)

Yadav, Rohit Kumar (yada6101@vandals.uidaho.edu)
Masters Computer Science, University of Idaho

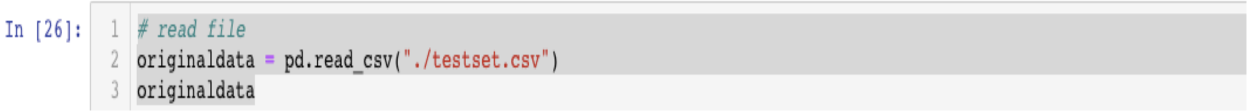
Implementation & Approach Description

Importing the Data and libraries:

Data can originate from myriad sources, and needs to be accumulated before it can be put to use. This can be done by directly importing files that may already be available in .csv or .xlsx formats. This could also be done by manually entering data tuples into a data repository, while referring to different sites. This can be a very tedious and time consuming process.

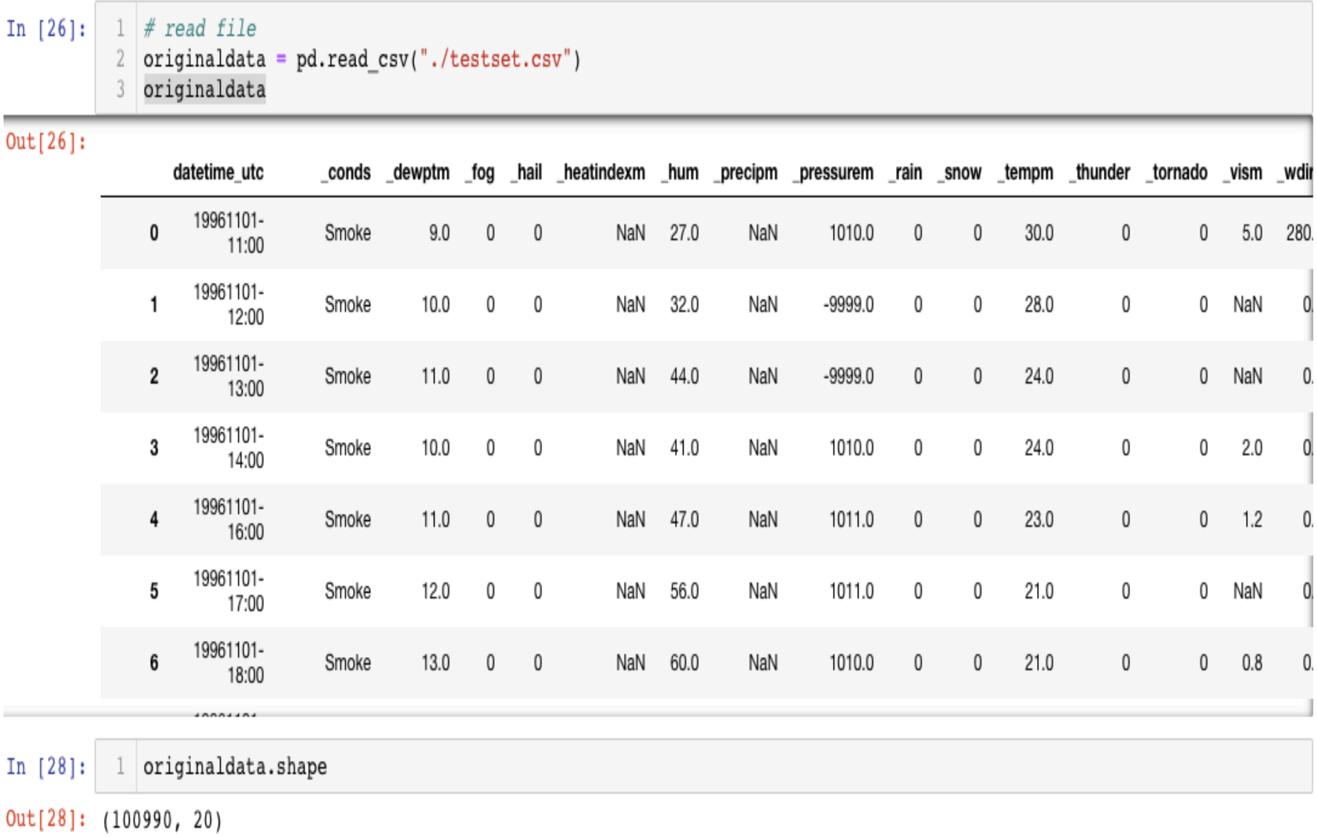


```
In [22]:
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import scipy as sp
import string
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.decomposition import PCA
from sklearn.linear_model import LinearRegression
from sklearn.linear_model import Ridge
from sklearn.neighbors import KNeighborsRegressor
from sklearn.ensemble import GradientBoostingRegressor
from sklearn.tree import DecisionTreeRegressor
import time
```



```
In [26]:
# read file
originaldata = pd.read_csv("./testset.csv")
originaldata
```

Yadav, Rohit Kumar (yada6101@vandals.uidaho.edu)
Masters Computer Science, University of Idaho



In [26]:

```
# read file
originaldata = pd.read_csv("./testset.csv")
originaldata
```

Out[26]:

	datetime_utc	_conds	_dewptm	_fog	_hail	_heatindexm	_hum	_precipm	_pressurem	_rain	_snow	_tempm	_thunder	_tornado	_vism	_wdir
0	19961101-11:00	Smoke	9.0	0	0	NaN	27.0	NaN	1010.0	0	0	30.0	0	0	5.0	280.
1	19961101-12:00	Smoke	10.0	0	0	NaN	32.0	NaN	-9999.0	0	0	28.0	0	0	NaN	0.
2	19961101-13:00	Smoke	11.0	0	0	NaN	44.0	NaN	-9999.0	0	0	24.0	0	0	NaN	0.
3	19961101-14:00	Smoke	10.0	0	0	NaN	41.0	NaN	1010.0	0	0	24.0	0	0	2.0	0.
4	19961101-16:00	Smoke	11.0	0	0	NaN	47.0	NaN	1011.0	0	0	23.0	0	0	1.2	0.
5	19961101-17:00	Smoke	12.0	0	0	NaN	56.0	NaN	1011.0	0	0	21.0	0	0	NaN	0.
6	19961101-18:00	Smoke	13.0	0	0	NaN	60.0	NaN	1010.0	0	0	21.0	0	0	0.8	0.

In [28]:

```
originaldata.shape
```

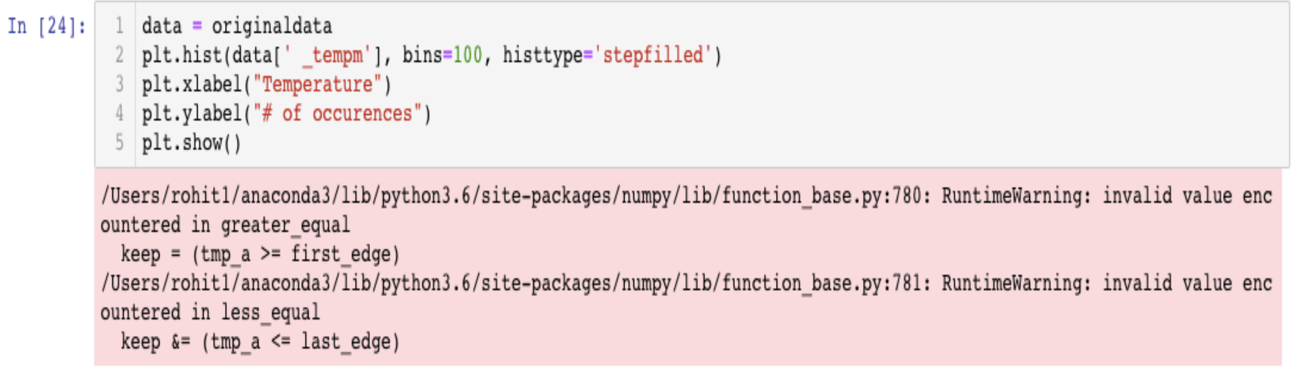
Out[28]:

```
(100990, 20)
```

Preliminary Visualization:

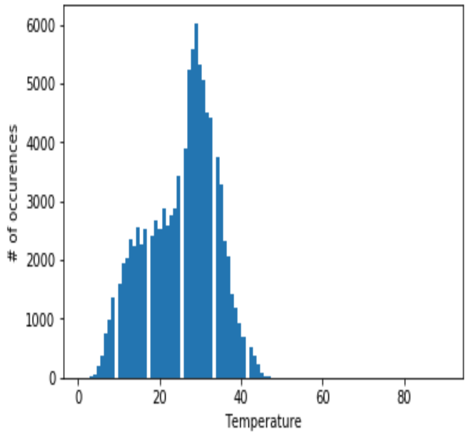
We are doing preliminary visualization to learn the data better and visualize them to best fit in our models this step basically analyze the data better and make the dataset ready for our different models and clustering

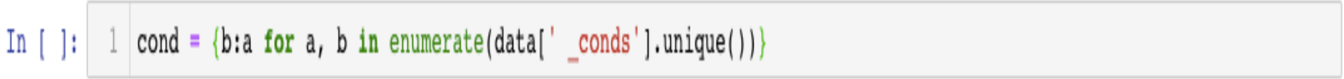
Yadav, Rohit Kumar (yada6101@vandals.uidaho.edu)
Masters Computer Science, University of Idaho



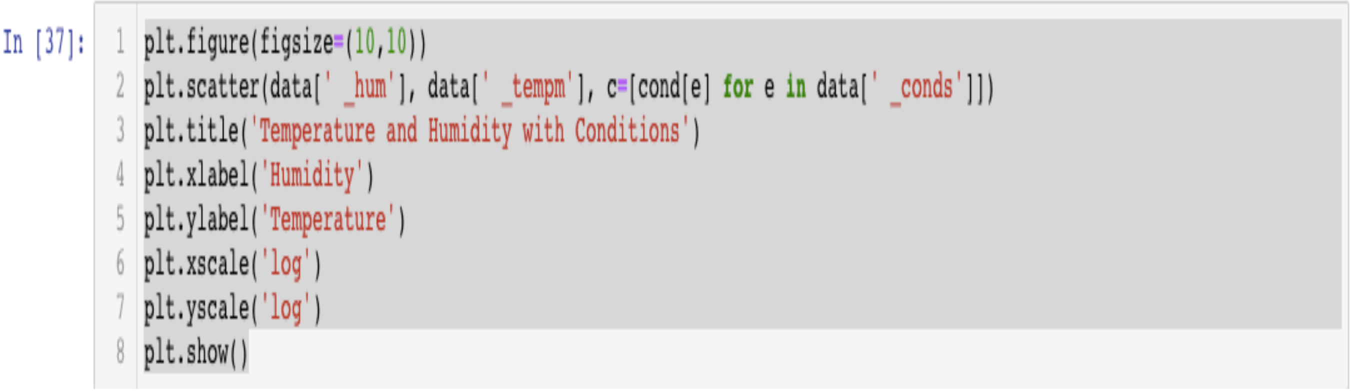
In [24]:
```
data = originaldata
plt.hist(data[' _tempm'], bins=100, histtype='stepfilled')
plt.xlabel("Temperature")
plt.ylabel("# of occurences")
plt.show()
```

```
/Users/rohit1/anaconda3/lib/python3.6/site-packages/numpy/lib/function_base.py:780: RuntimeWarning: invalid value enc
ountered in greater_equal
  keep = (tmp_a >= first_edge)
/Users/rohit1/anaconda3/lib/python3.6/site-packages/numpy/lib/function_base.py:781: RuntimeWarning: invalid value enc
ountered in less_equal
  keep &= (tmp_a <= last_edge)
```





In [ ]:
```
cond = {b:a for a, b in enumerate(data[' _conds'].unique())}
```



In [37]:
```
plt.figure(figsize=(10,10))
plt.scatter(data[' _hum'], data[' _tempm'], c=[cond[e] for e in data[' _conds']])
plt.title('Temperature and Humidity with Conditions')
plt.xlabel('Humidity')
plt.ylabel('Temperature')
plt.xscale('log')
plt.yscale('log')
plt.show()
```

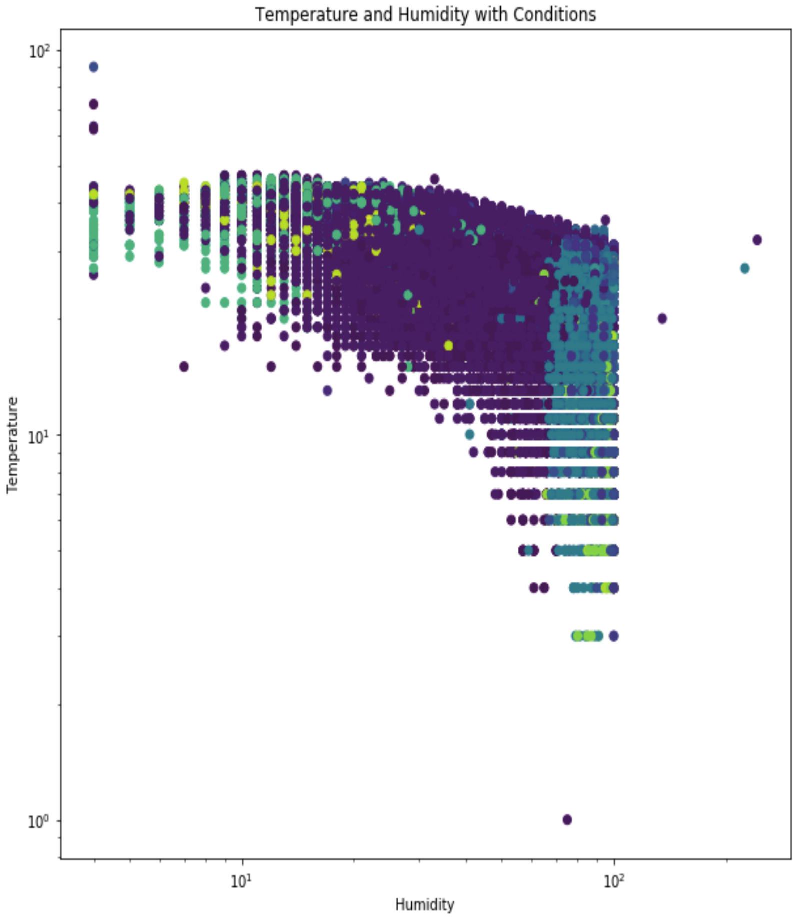
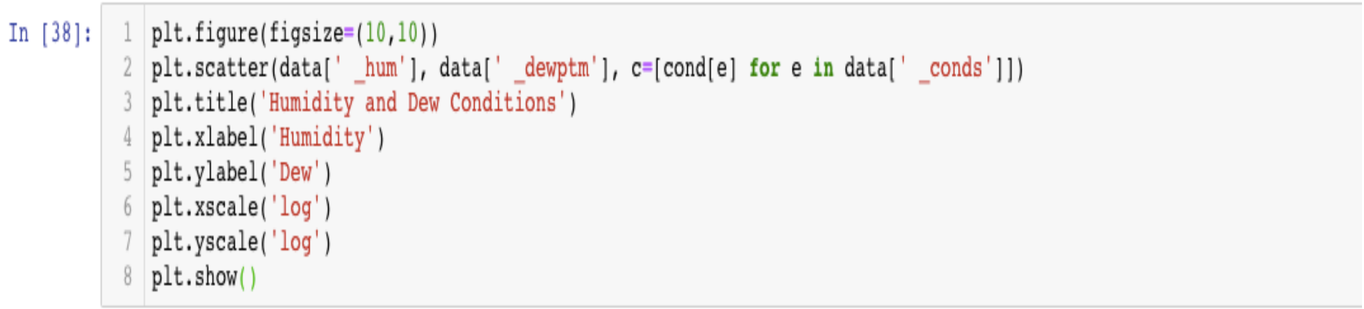


Fig: Humidity vs Temperature graph



```
In [38]: plt.figure(figsize=(10,10))
         plt.scatter(data[' _hum'], data[' _dewptm'], c=[cond[e] for e in data[' _conds']])
         plt.title('Humidity and Dew Conditions')
         plt.xlabel('Humidity')
         plt.ylabel('Dew')
         plt.xscale('log')
         plt.yscale('log')
         plt.show()
```

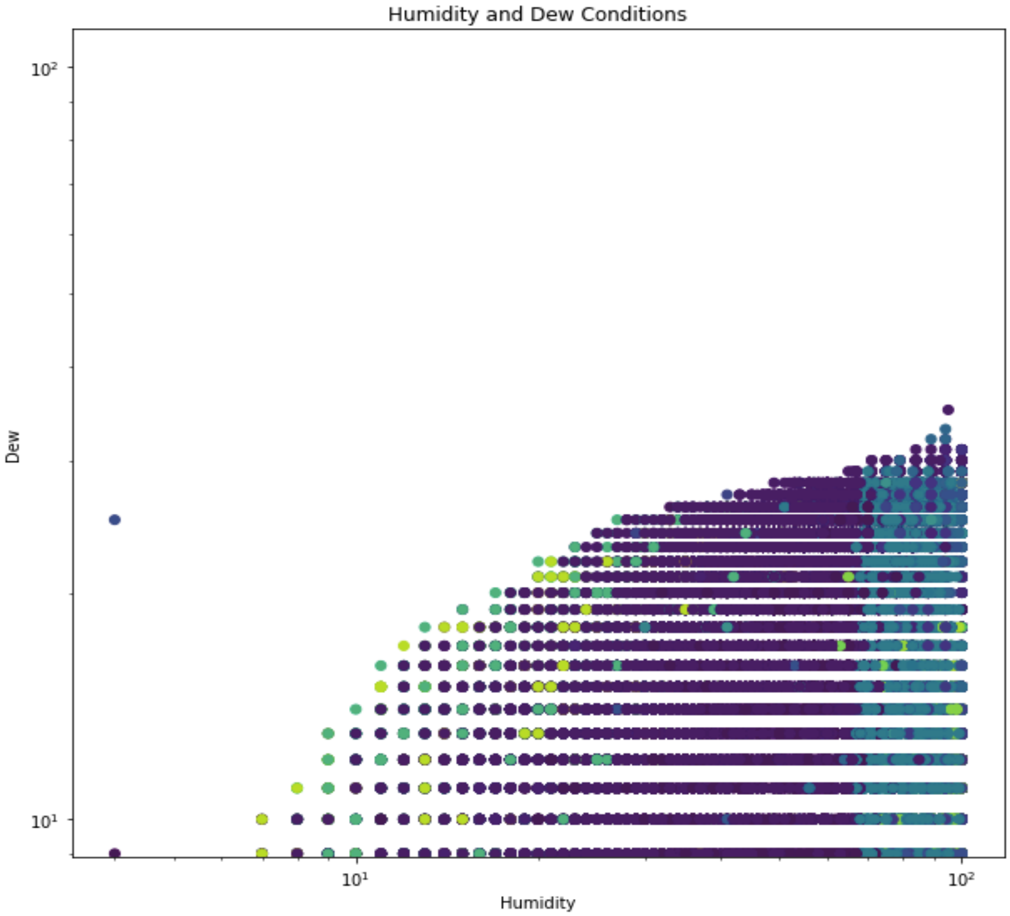


Fig: Humidity vs Dew graph

The upper two graphs are the preliminary visualizations of the data by analysis of temperatuere-humidity and dew-humidity to find the correlation between them for our models and clustering analysis.

In temperature-humidity correlation we found the negative correlation between them which is inversely related to each other where If the temperature is high humidity is low because its dries the water in the whole region and when humidity is high temperature gets low which means It gets colder and wet region.

Data Preprocessing

This stage involves assimilating all the datasets from the Delhi-weather-data. We have preprocessed all the data and make it ready for the analysis and visualizations for our different models and clustering.

Datasets in any data mining applications can have missing data values. These missing values can get propagated due to faulty sensors, or lack of communication among the components in a data collection system. These missing values can affect the performance of a data mining system, and need to be addressed accordingly. This can be done by representing these missing values by an approximated mean value. In this we have many missing values of the variables _precipm and heatindexm are replaced with their corresponding averaged representatives.

Data Analysis

This stage involves various parts of analysis unsupervised and supervised learning.

Feature Set Description

We have performed various regression models and clustering for finding the temperature with similar climate conditions in the region.

Different regression models used

- Linear regression
- Ridge regression
- Decision Trees
- K-nearest neighbors
- Gradient Boosted regression

Different methods used for clustering analysis:

I will perform different clustering methods for comparing the similar climatic conditions in the region

- Kmeans
- Ward hierarchical clustering
- Gaussian Mixtures
- Aglomerrative Clustering

System Analysis:

Functional requirements for our project is listed below:

- Rain Predictions- we gonna predict the rain conditions in the region by comparing the similar climatic conditions.
- Humidity Predictions- by using regression we can predict the humidity condition in the region by comparing with any variables like temperature or season in our dataset.
- Season-wise predictions- through the visualization of clustering and serialtion analysis we can predict the seasons in different regions.
- Temperature predictions- we can find the different patterns using the regression for predicting the temperature in different region for eventually improving the climate conditions of whole area.

Conclusions:

This project highlights the application of machine learning and data mining algorithms in the field of weather. Climate and temperature predictions, if presented in a proper format to the end-users, I believe the weather conditions can be controlled and also this will offer insights to the climate in the region.

References:

1. UNHCR - The UN Refugee Agency. http://www.unhcr.org/en-us/news/latest/2008/12/493e9bd94/top-unhcr-official-warns-displacement-climate-change.html. [Ac- cessed: 2017-04-10].

2. Fast Facts About Climate Change - National Wildlife Federation. https://www.nwf.org/Eco-Schools-USA/Become-an-Eco-School/Pathways/Climate-Change/Facts.aspx. [Accessed: 2017-04-04].

3. The Guardian - ’A tipping point’: record number of Americans see global warming as threat. hhttps://www.theguardian.com/environment/2016/mar/18/ climate-change-record-concern-us-global-warming-poll. [Accessed: 2015-04-10].