**Week five: Data Science Capstone Project:**

**Introduction**

This section highlights the tittle of the project and problem statement.

**Project Name**

Analysis of USA Weather Anomalies.

**Problem Statement**

The world is faced with climate change crisis than never before, which is no longer a phenomenon of scientific research but today’s reality. This has been echoed in the recent UN general assembly and in world-wide demonstrations. Climate change is a long-term shift in weather conditions identified by changes in temperature, precipitation, winds, and other indicators. Climate change can involve both changes in average conditions and changes in variability, including, for example, extreme events. Human activity is the main cause of climate change. People burn fossil fuels and convert land from forests to agriculture. Since the beginning of the Industrial Revolution, people have burned more and more fossil fuels and changed vast areas of land from forests to farmland. Burning fossil fuels produces carbon dioxide, a greenhouse gas. It is called a greenhouse gas because it produces a “greenhouse effect”. The greenhouse effect makes the earth warmer, just as a greenhouse is warmer than its surroundings. Carbon dioxide is the main cause of human-induced climate change.t stays in the atmosphere for a very long time. Other greenhouse gases, such as nitrous oxide, stay in the atmosphere for a long time. Other substances only produce short-term effects. Not all substances produce warming. Some, like certain aerosols, can produce cooling. USA is the worst climate polluter according to data from Carbon Brief. Through its current president, it withdrew from the global pact to reduce carbon emissions despite being the number one polluter. This project therefore aims to study weather changes in USA to establish the magnitude of carbon emissions on temperature and precipitation, which will enable current and future stokeholds to open their eyes and reverse the crisis.

**Data Description**

This section shows the data that will be used in the analysis with its source

More than 3 million records of historical US data temperature outliers by date and latitude/longtitude coordinates.

Temperature outliers from 1964-2013. Reporting station ID, name, min/max temperature, as well as degree coordinates of the recorded weather. Original weather data collected from NOAA.

Each entry represents a report from a weather station with high or low temperatures that were historical outliers within that month, averaged over time. Note: This table's columns contain data that was collected from NOAA as well as data that was calculated using Enigma's assortment of weather data. The direct source of the information is identified in the description of the column.

**Feature set**

The data includes the following features:

|  |  |
| --- | --- |
| Feature label | Feature Description |
| date\_str | Date of temperature recording |
| degrees\_from\_mean | Degrees from the mean temperature(temperature outliers) |
| id | Unique Identification of each temperature |
| longitude | Longitude of the station name |
| latitude | Latitude of the station name |
| max\_temp | Maximum temperature recorded at station name on a given date |
| min\_temp | Minimum temperature recorded at station name on a given date |
| station\_name | Name of the station |
| type | Weather description |
| serialid | Cummulative identification of each temperature |

**Data source**

The data set can readily be accessed from the URL below;

https://data.world/carlvlewis/u-s-weather-outliers-1964/file/weather-anomalies-1964-2013.csv

**Methodology**

This section represents the main component of the report,which include:data preparation and pre-processing, exploratory data analysis , inferential statistical testing and machine learnings used .

Data loading, preparation and pre-processing:

Data loading , preparation and pre-processing is achieved with the help of pandas library.

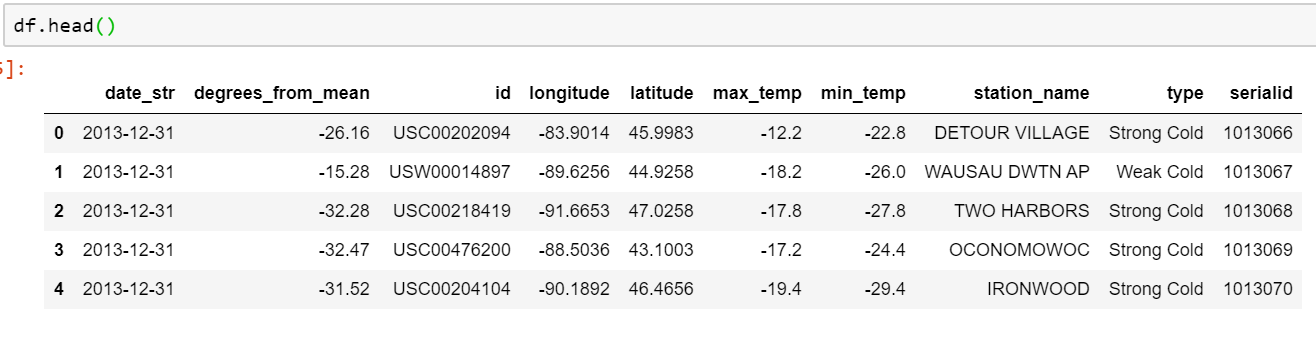
Data Loading:

Data is stored and manipulated in memory using numerical python(numpy) and Pandas libraries. The dependencies are loaded using import numpy as np and pandas as pd. Pd.read\_cvs(file\_name) function is used to load data into memory.

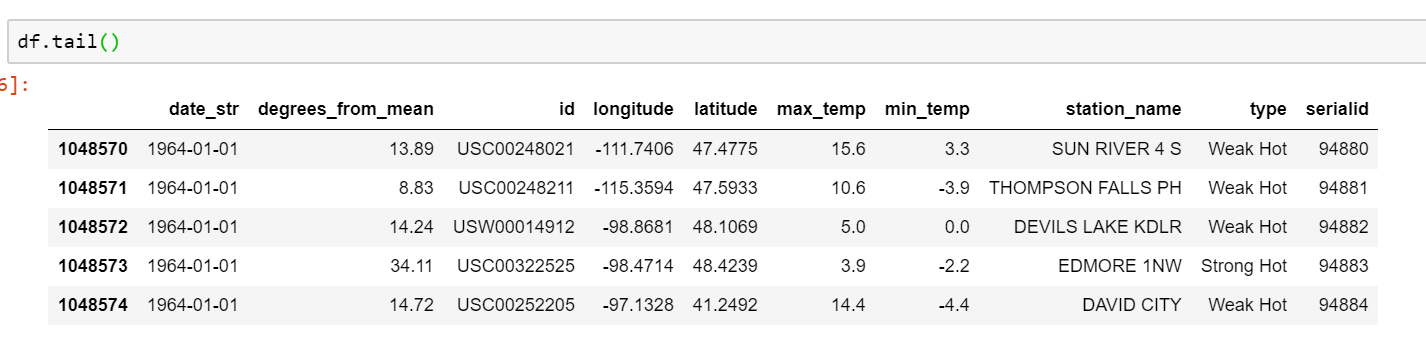
preparation and pre-processing:

Data is not always ready for analysis. Quick view to the nature of the dataset is achieved using pd.head() for first five rows,pd.tail() for last five rows, and pd.info() for the general information for the variables as shown below:

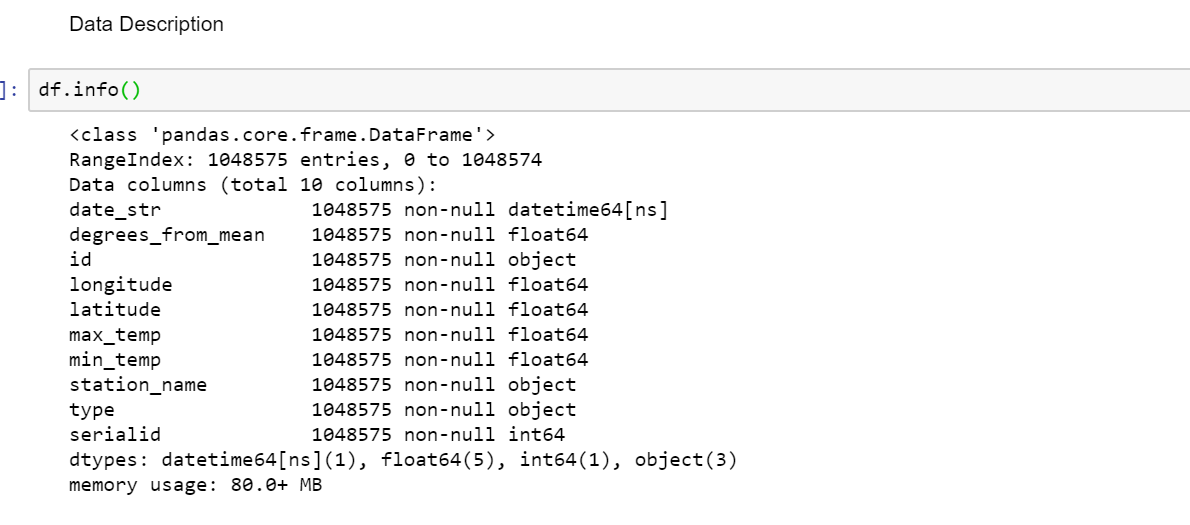
Five first rows:



Five last rows:



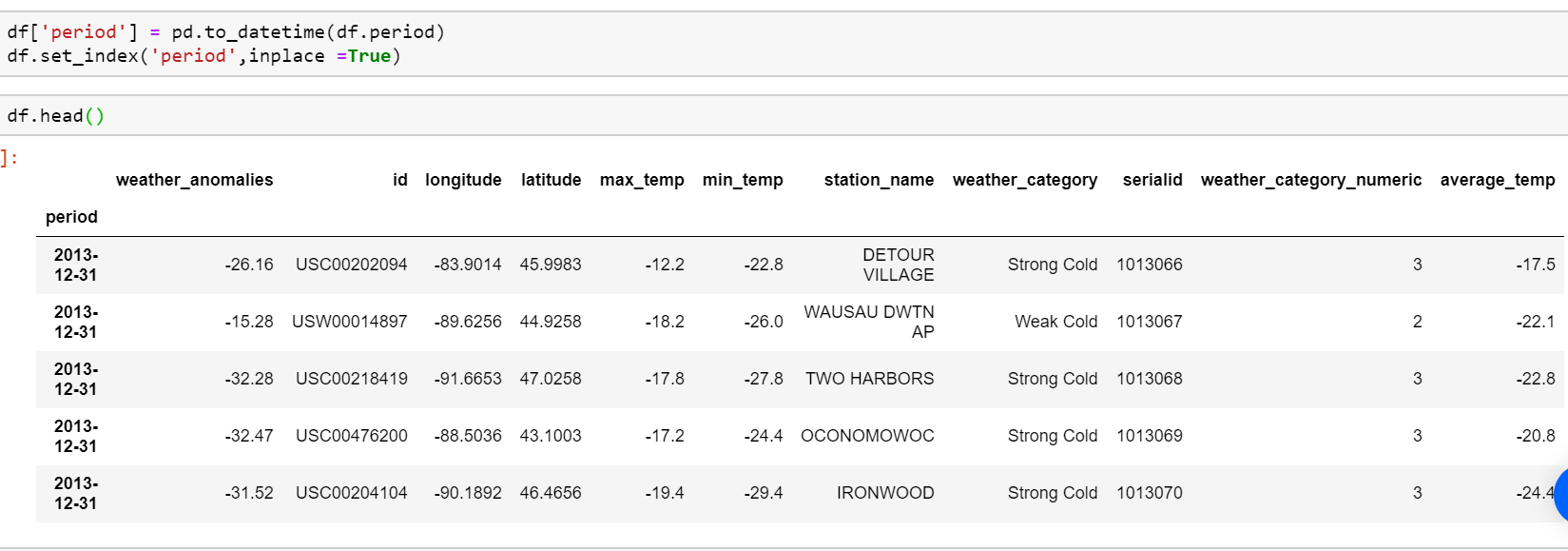
Data Type description:



Changing the weather category feature into numeric categories

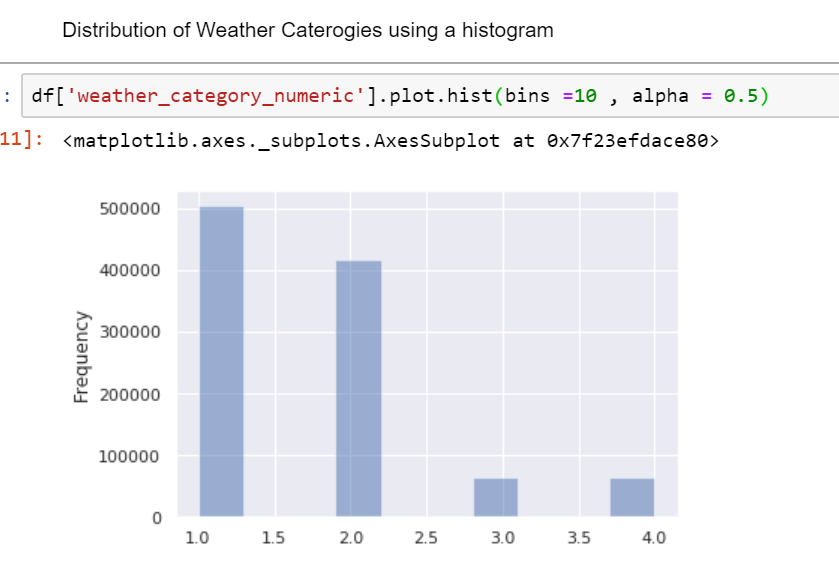


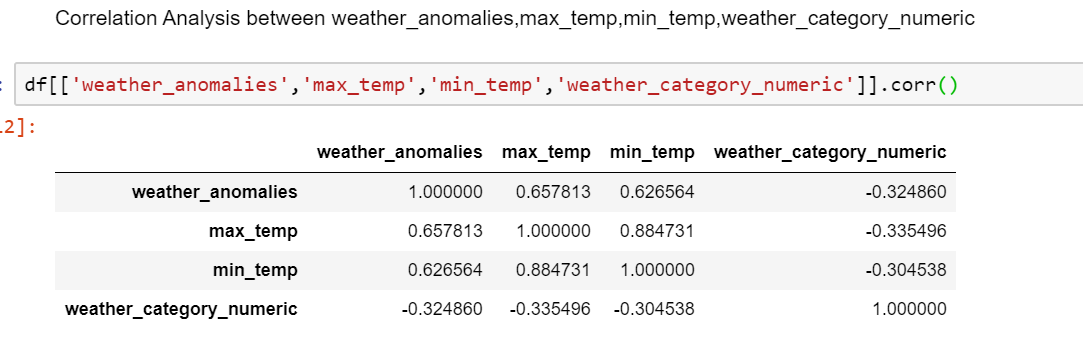
Changing the period feature to datetime data type and setting it as index



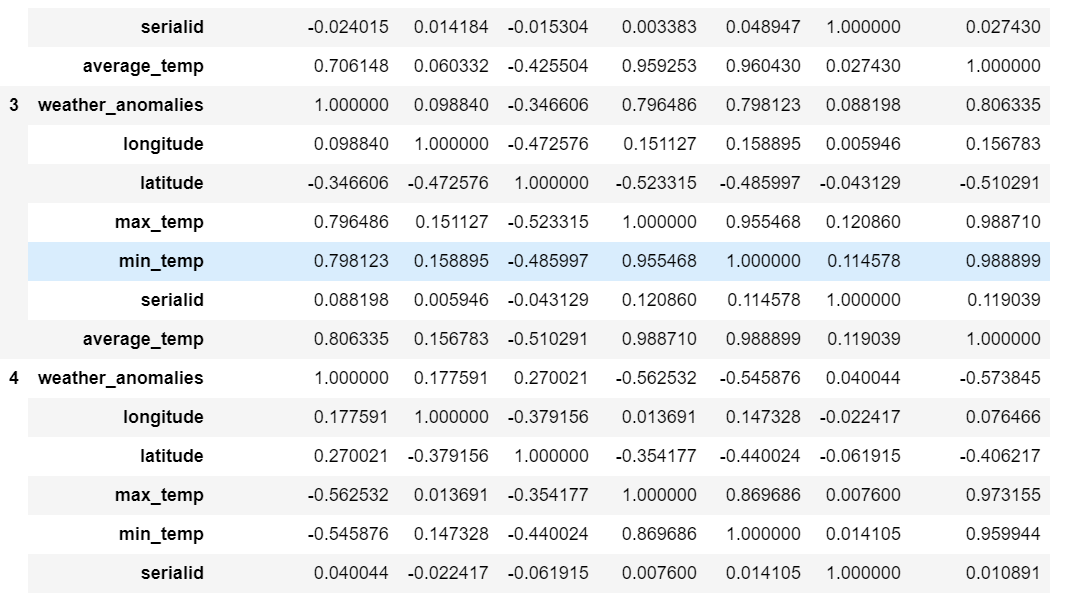
**Exploratory Data Analysis:**

This presents any descriptive analysis performed so as to determine the distribution of data. This include:Graphics like histogram, Correlation analysis, Visual Time series Analysis and table summaries.

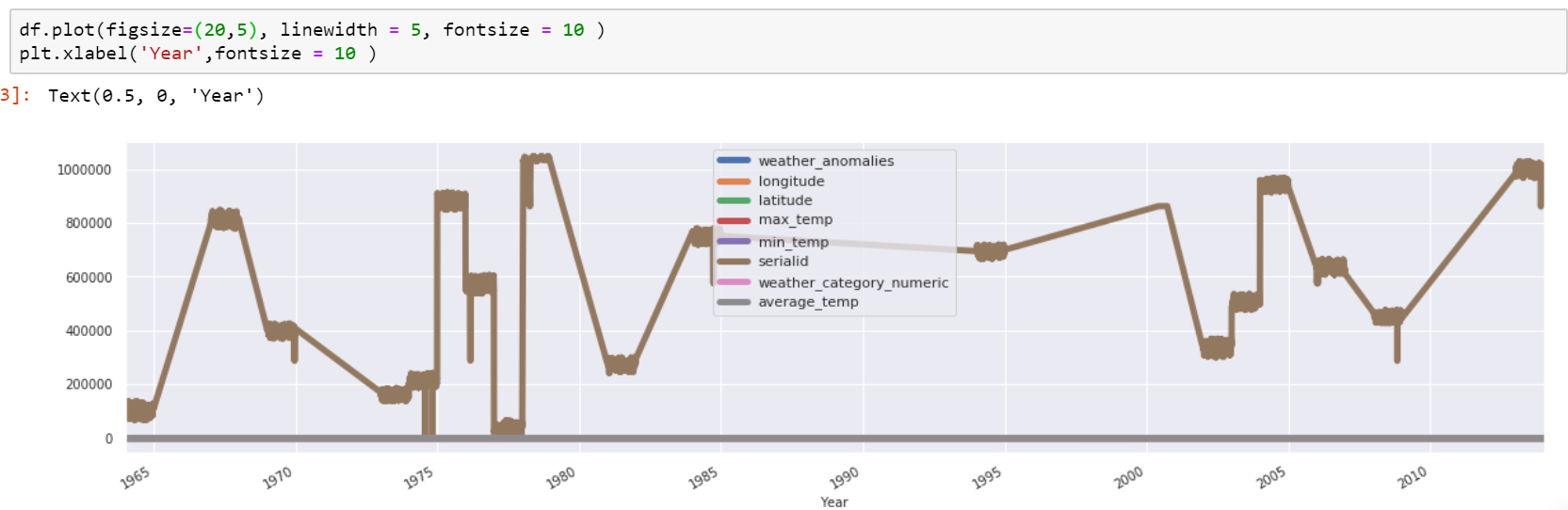


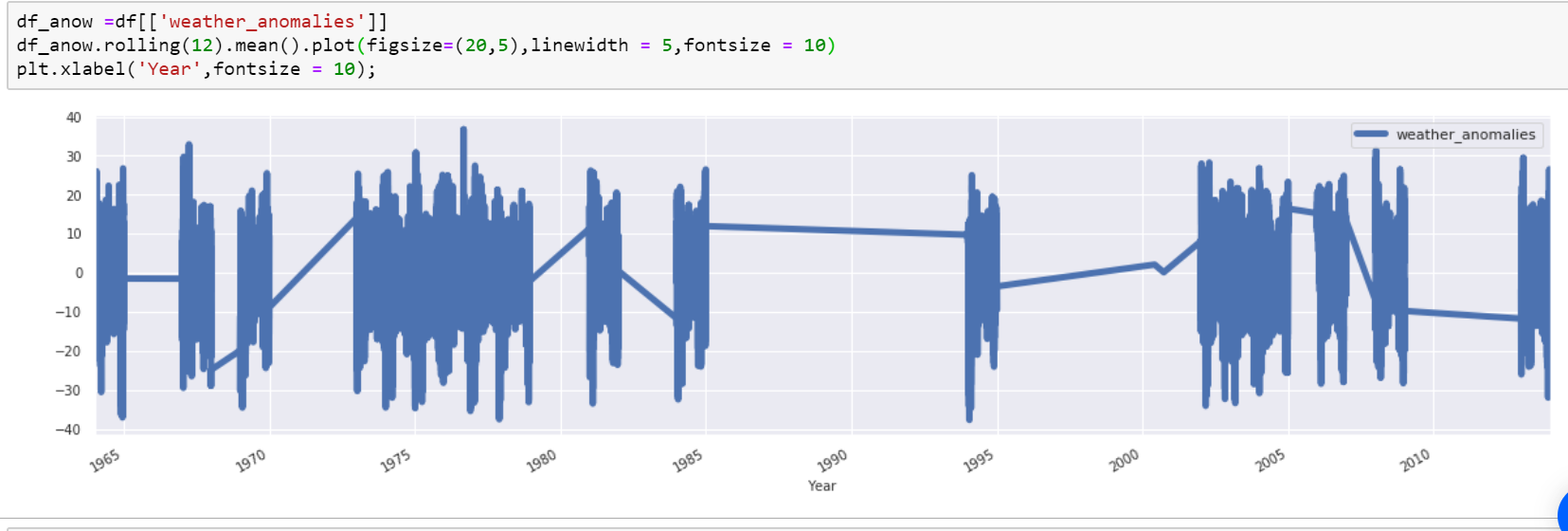


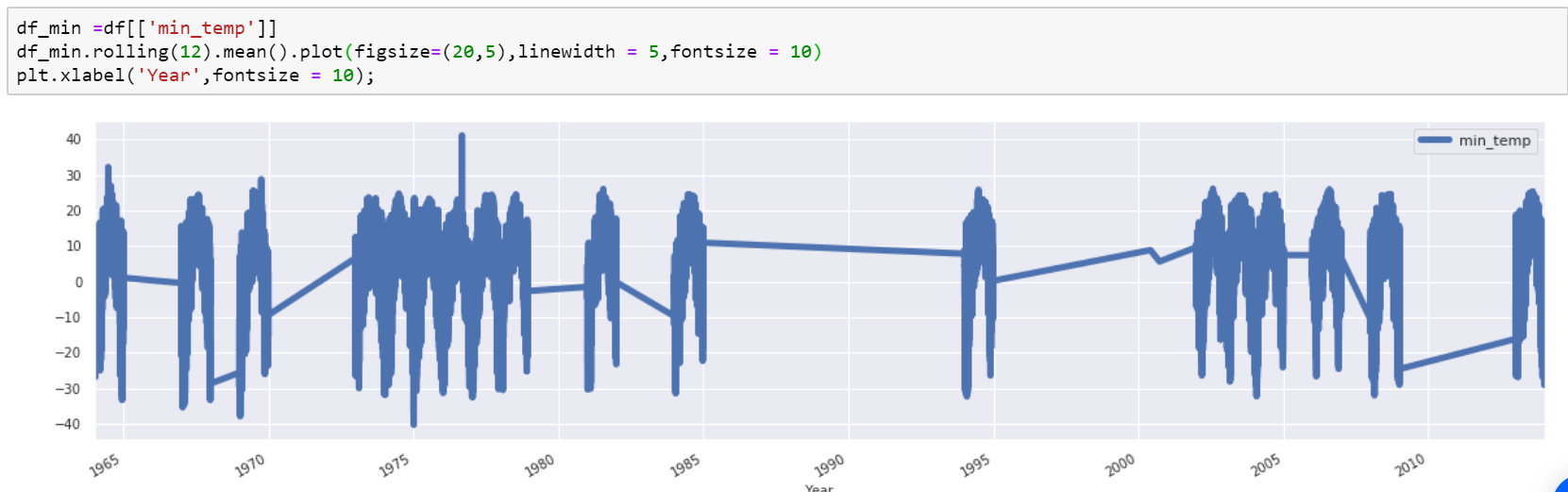


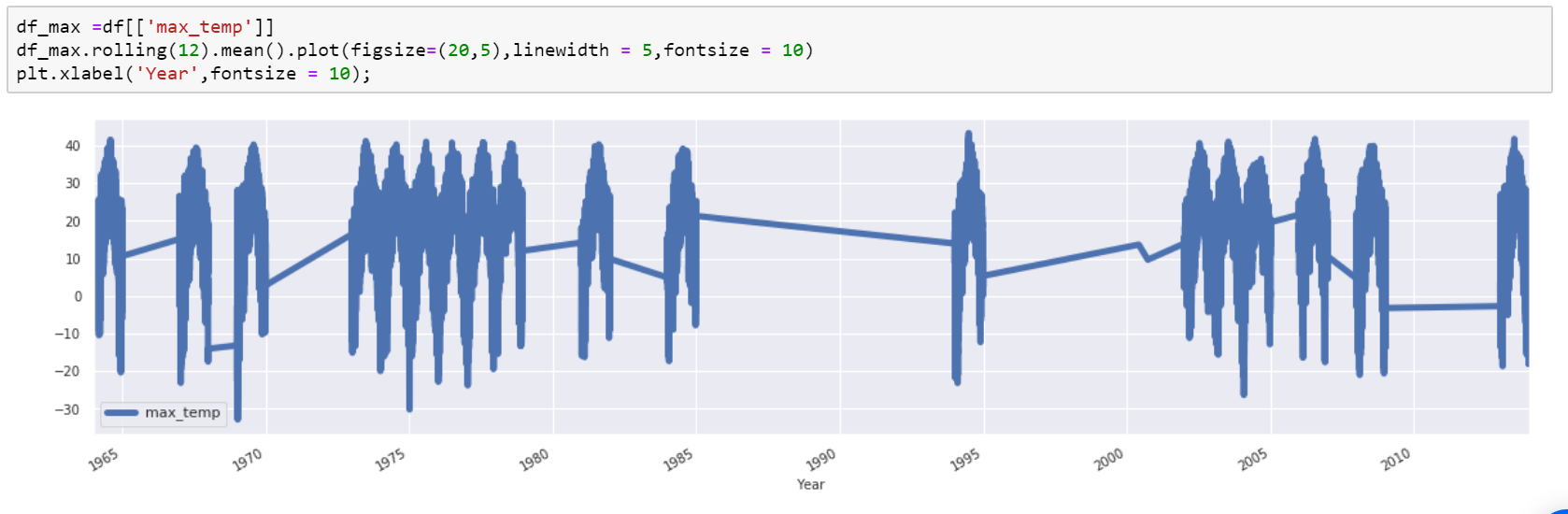


Visual Time Series Analysis:

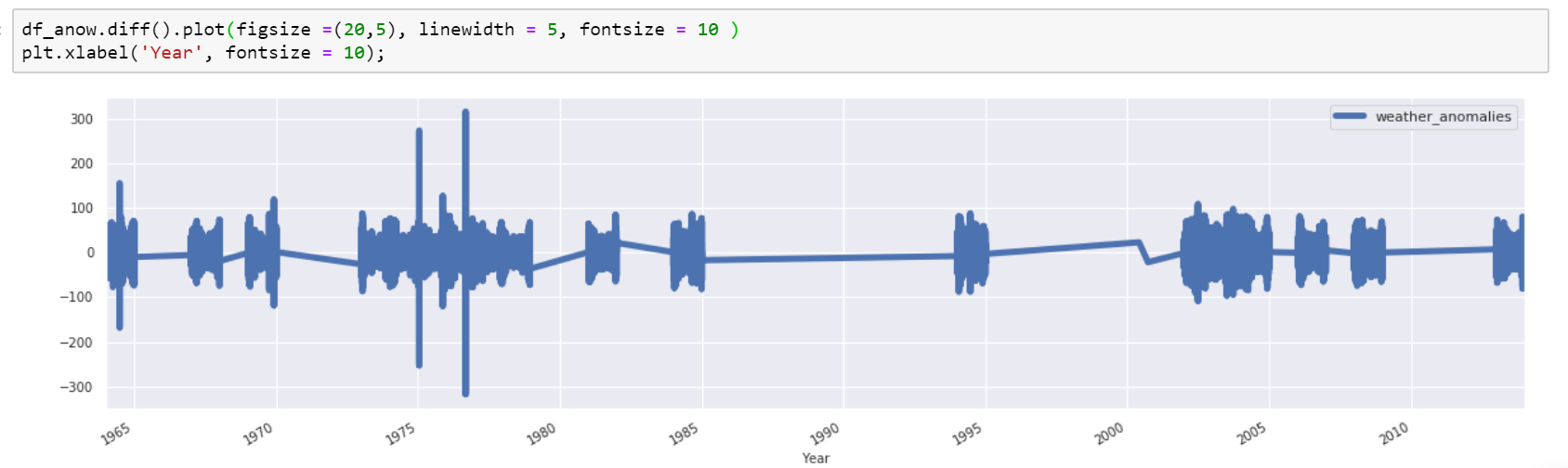


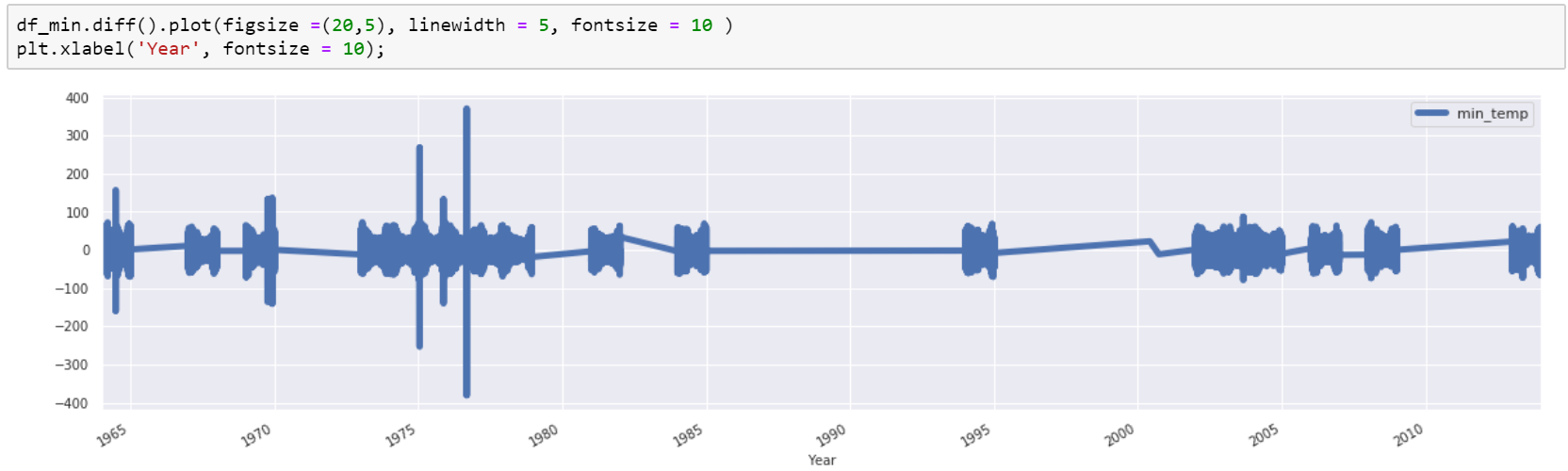


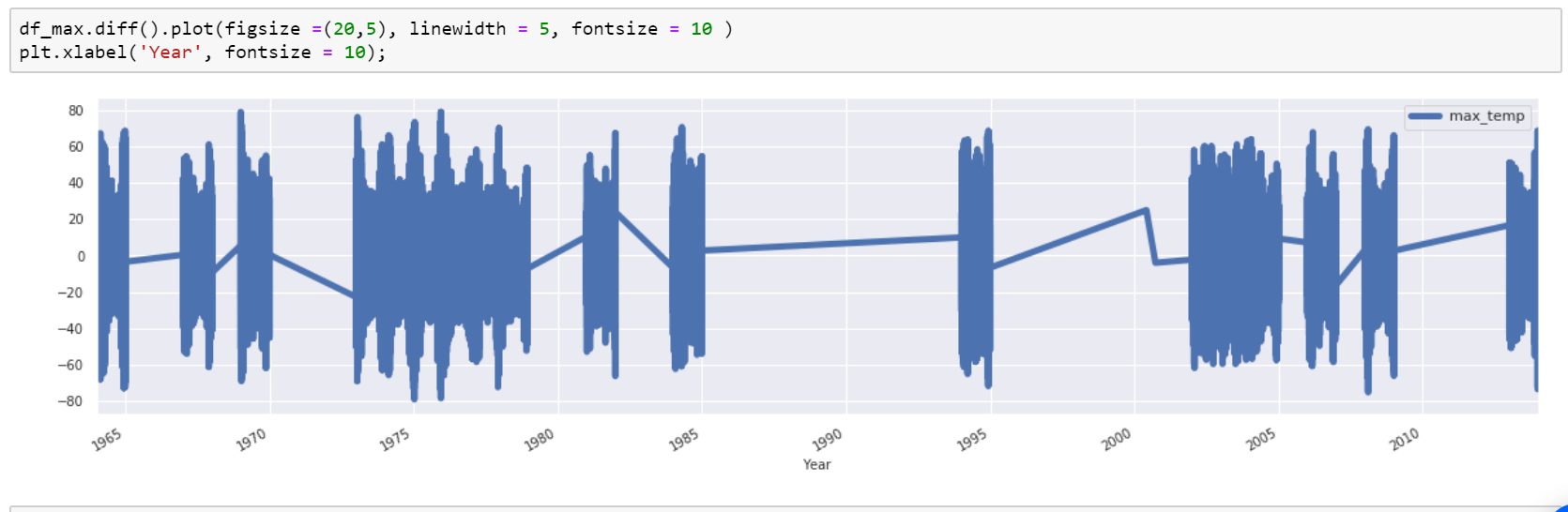


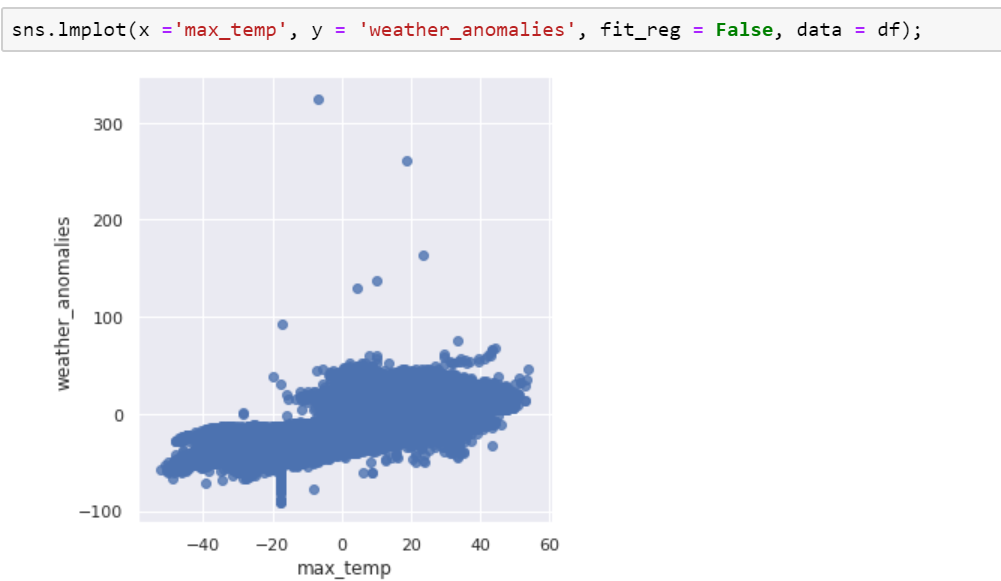








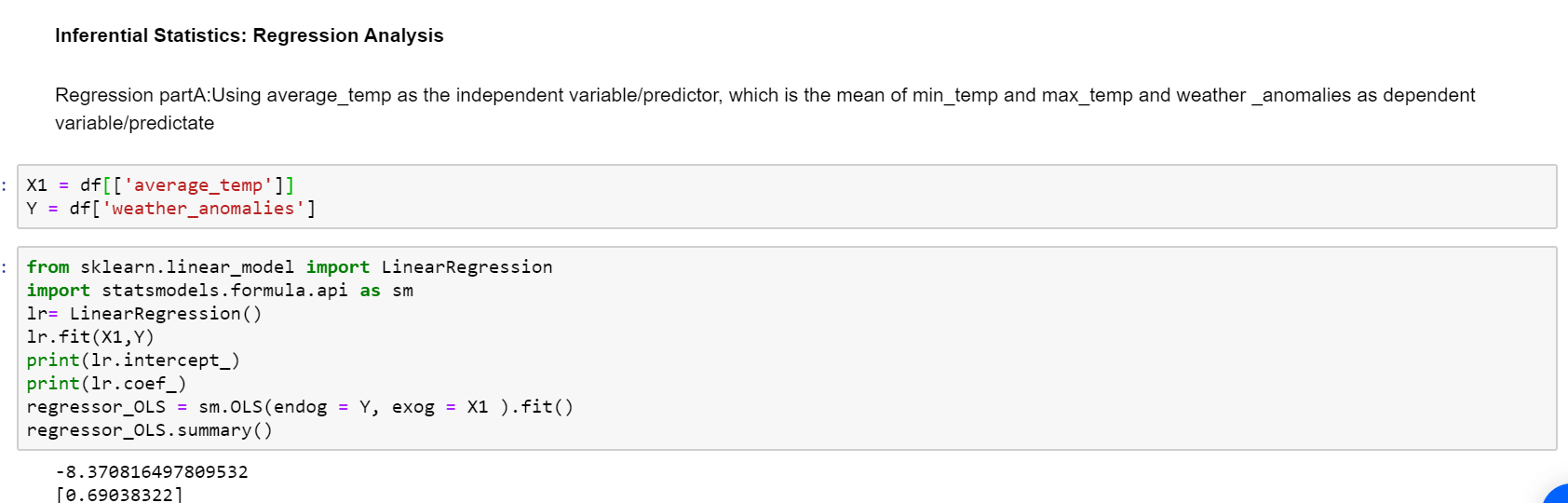




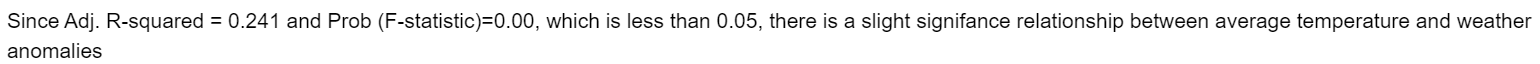
Inferential statistical test:

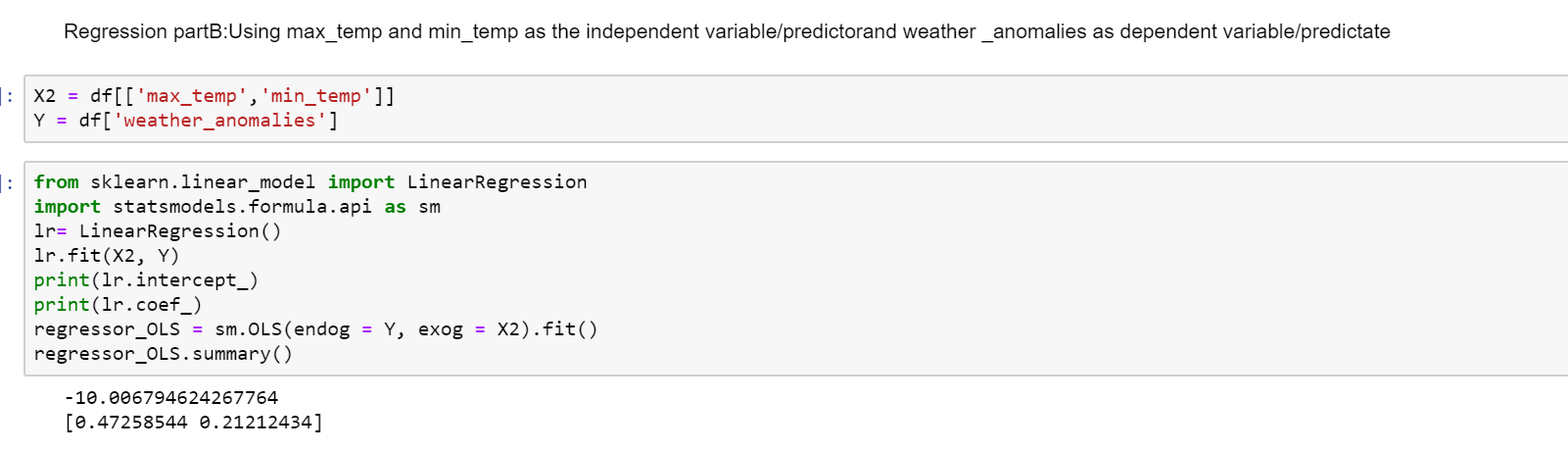
This presents any hypothesis testing performed. This was achieved using Regression Analysis and analysis of Variance.Regression analysis was used to determine the relationship between weather anomalies, min and max temperatures and different weather categories.

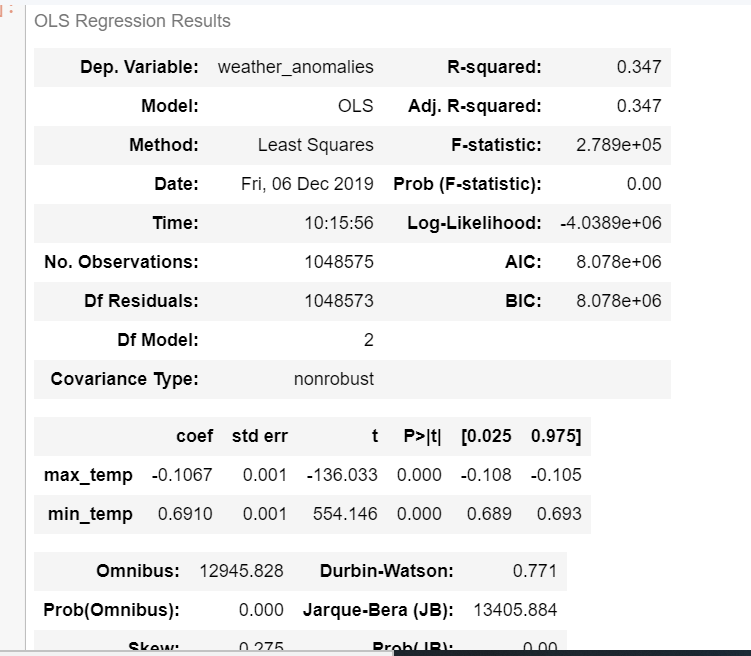
Analysis of variance(ANOVA) was used to determine the relationship between weather anomalies and the different weather categories.

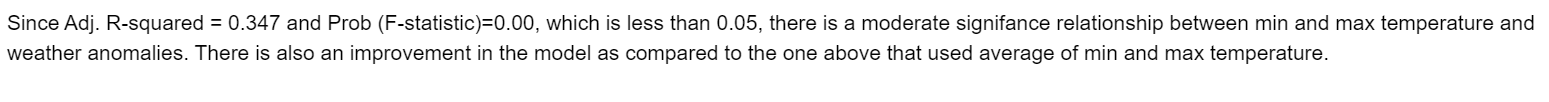


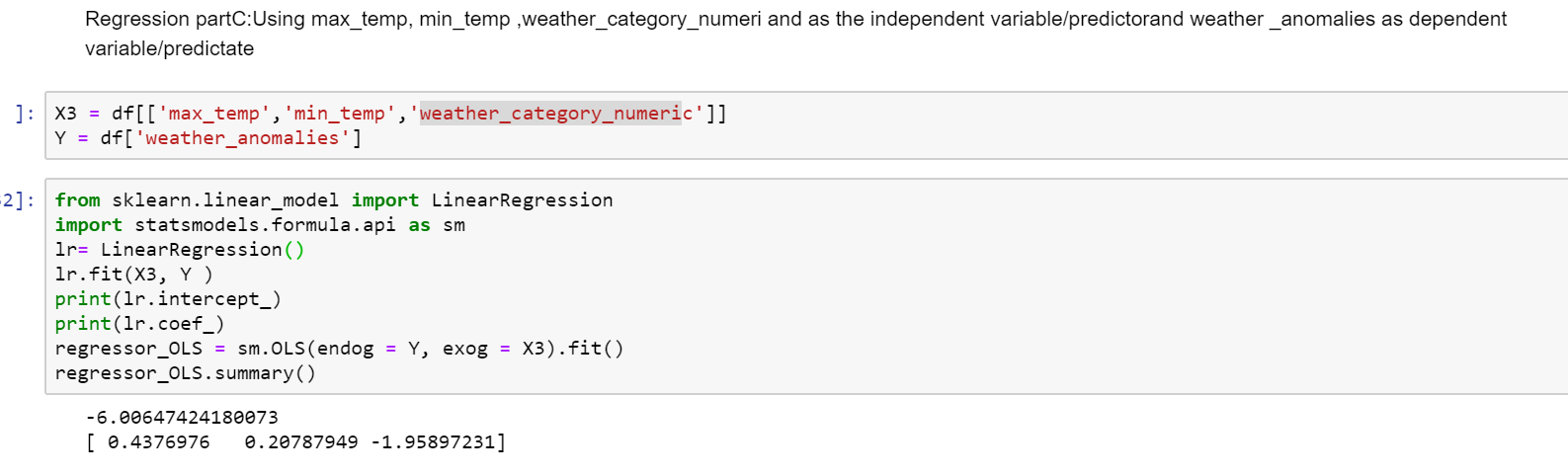


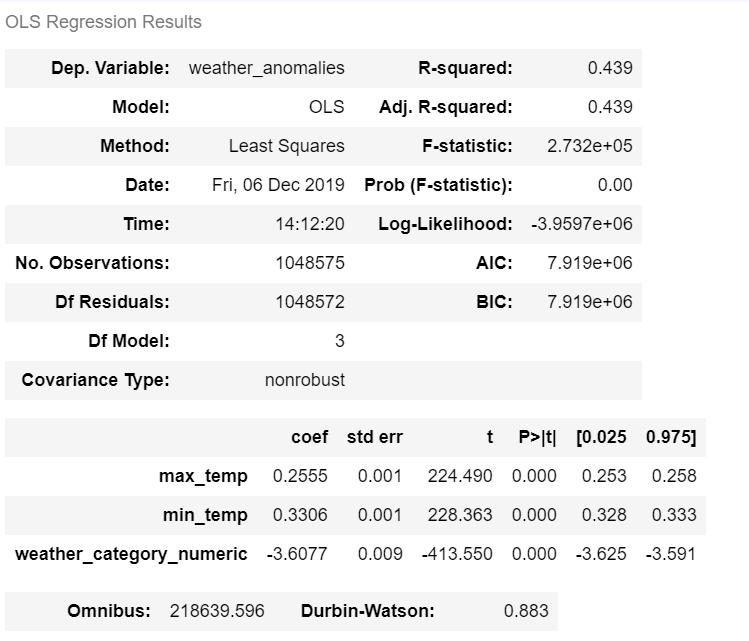


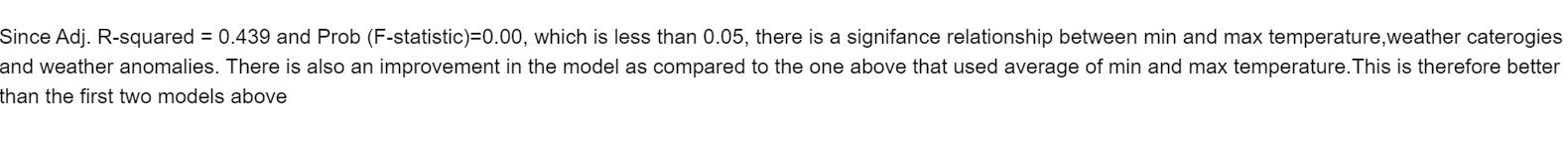


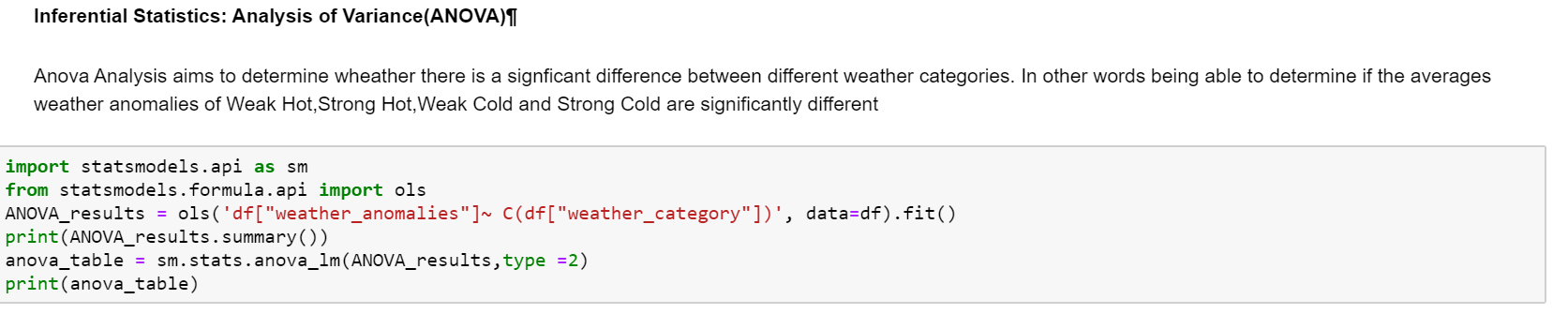


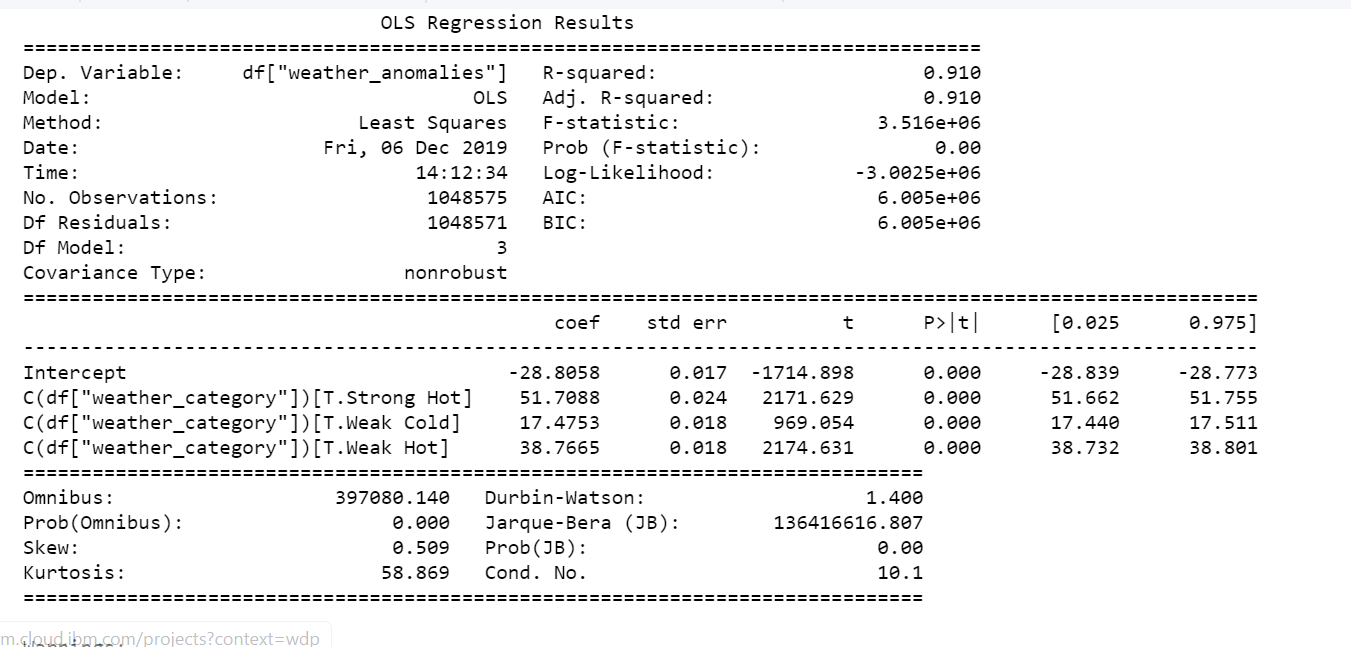


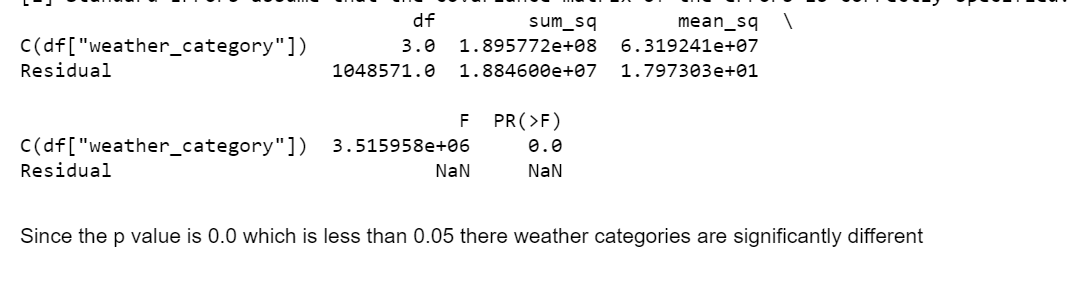








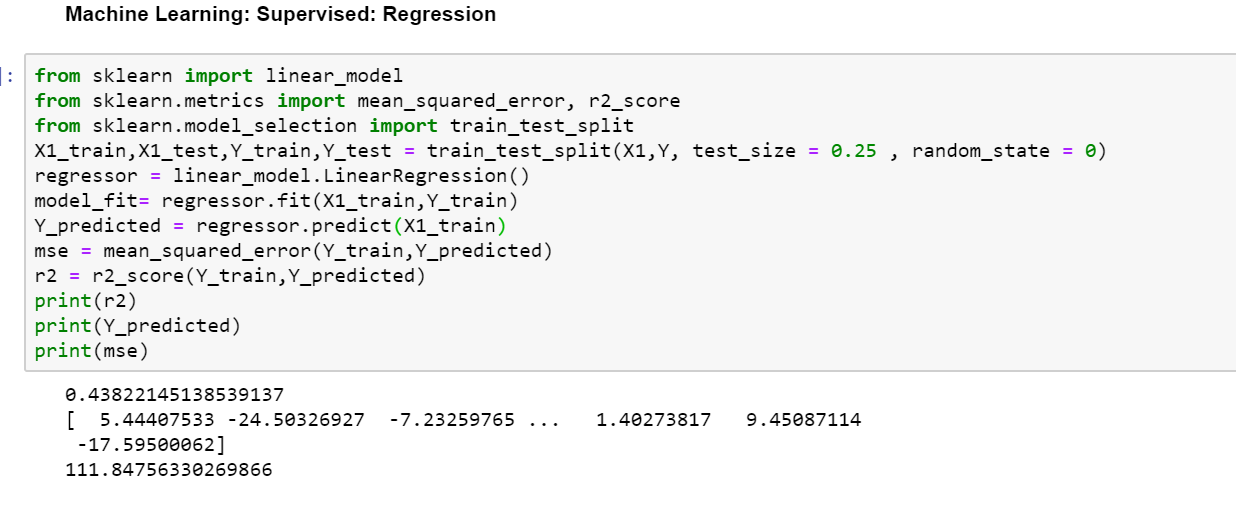


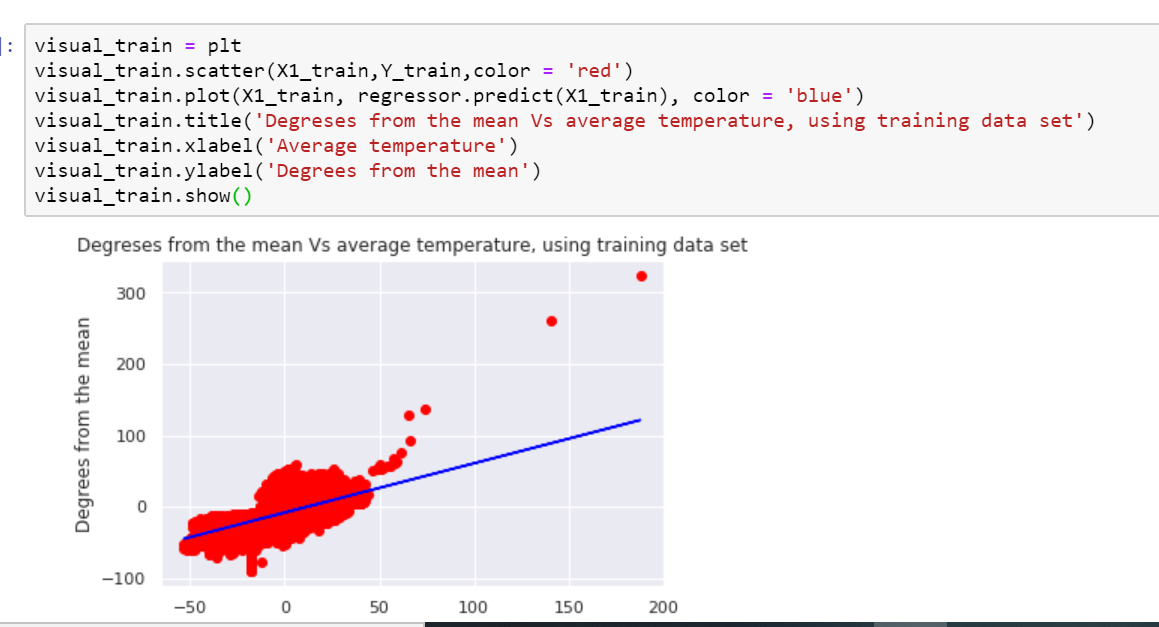


**Machine learning techiniques were also used for prediction:**

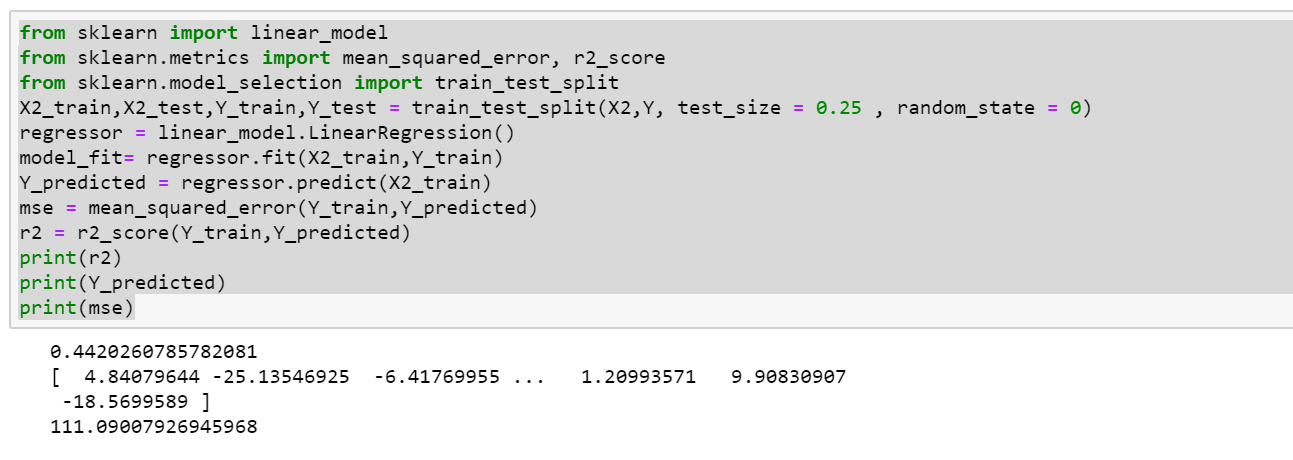
Liner Regression was used to predict weather anomalies using min,max temperatues and weather categories as the predictor variables. This was achieved in three steps: partA, which predicts weather anomalies using average temperature. PartB, which predicts weather anomalies using both min and max temperatures, and PartC, which predicts weather anomalies using both min and max temperatures ,and weather categories.

Part A: predicts weather anomalies using average temperature.

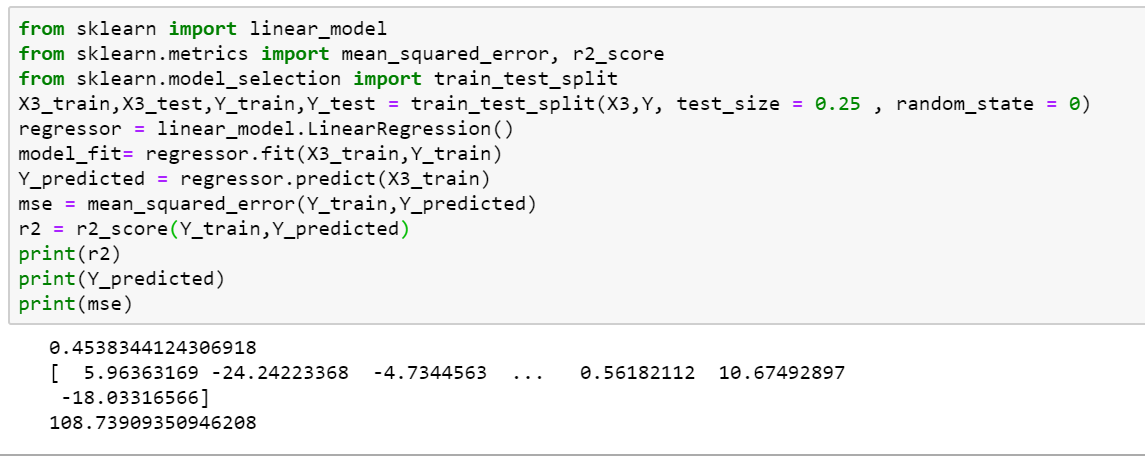




**PartB:** predicts weather anomalies using both min and max temperature.

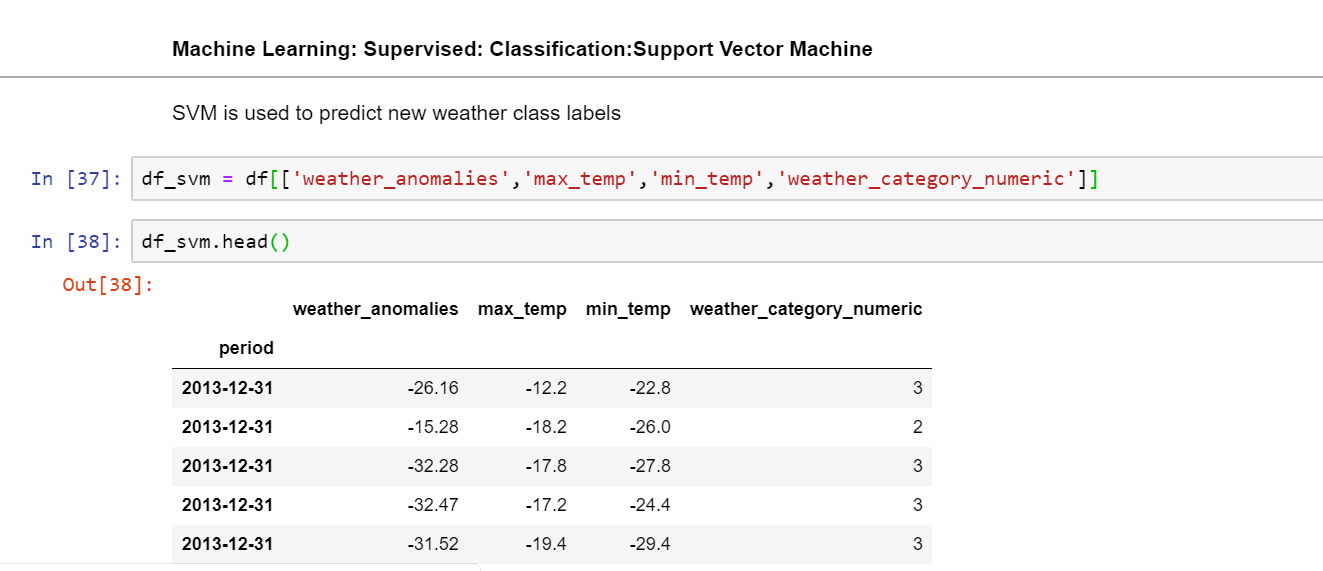


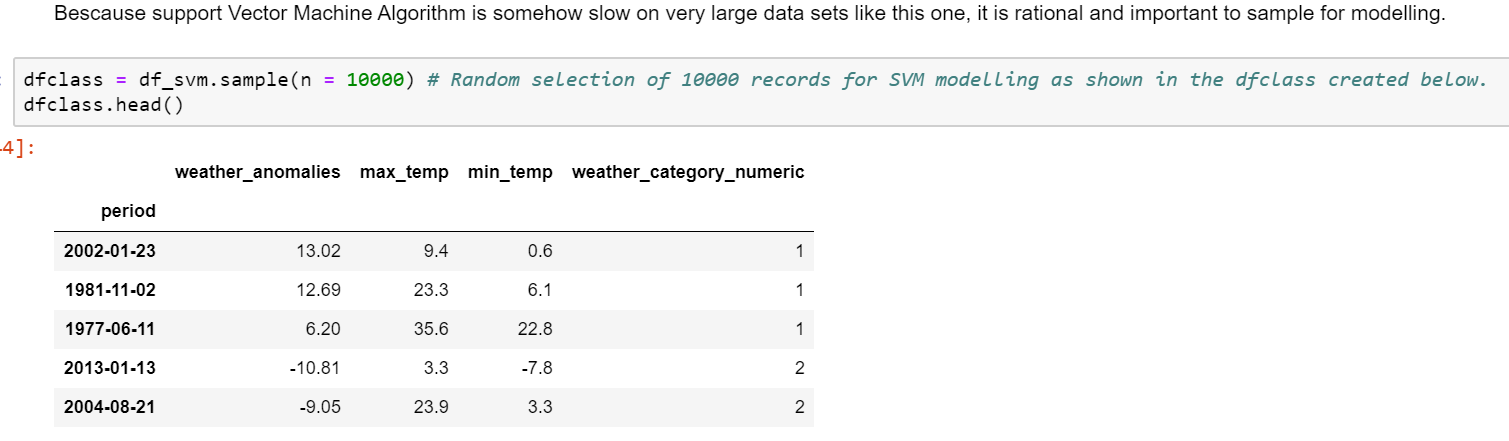
**PartC:** predicts weather anomalies using both min and max temperature, and weather categories.

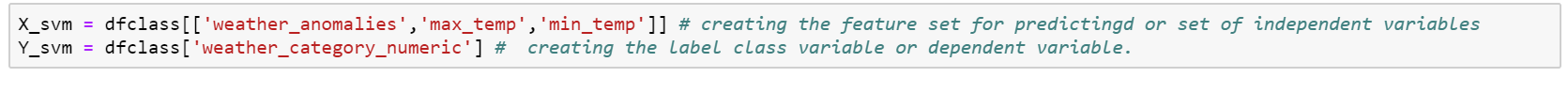


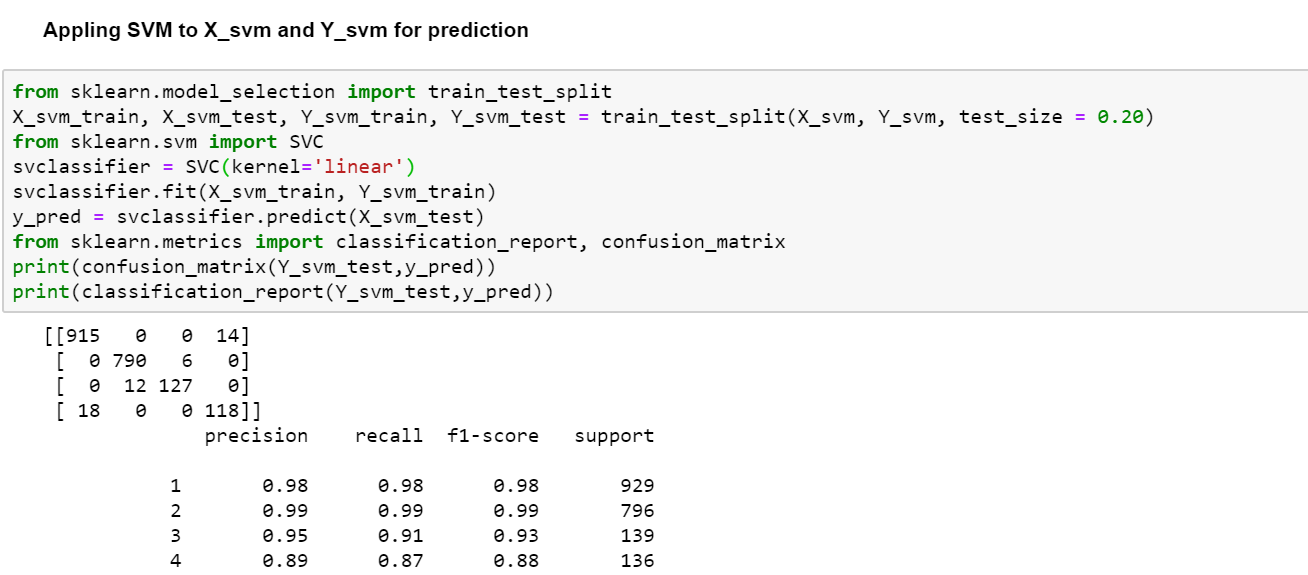
Machine learning classification was achieved using Support Vector Machine.

A ramdom sample of 10000 was used in the process.











**Conclusion:**

There is a significant relationship between weather anomalies and min, max and weather categories