Basic UFO Sightings Statistics

Brief Introduction:

In this project, this dataset has been used for exploration and finding some basic but interesting insights about UFO sightings around the world, from the data. Some visualizations have been done from perspective of length of UFO encounter timing, longitude of location of sighting, latitude of location of sighting and so on;

Key Work Areas:

1. Data Cleaning, Imputation, and Wrangling
2. Data Visualization

:Interpreting Columns of Dataset

Some columns of the dataset have been discarded as they didn't prove useful in the visualizations. The columns used in the visualizations are briefly described below:

1. **city –** city of the UFO sighting (actually it even has state and/or country for quite a few rows, due to improper recording of data)
2. **state/province –** code of province or state, of the country in which the sighting took place
3. **country** – country code of the country of sighting
4. **UFO_shape** – shape of the UFO as reported in the sighting in the data
5. **length_of_encounter_seconds** – length of the UFO sighting duration in seconds.
6. **date_documented** – reported date of the UFO sighting
7. **latitude –** latitude coordinate of the location of UFO sighting
8. **longitude** – longitude coordinate of the location of UFP sighting

:Work Done in Code

The Python script “main.py” generates 11 versions of the actual dataset, numbered from v0 (version 0) to 10. Each version is the output after having done some data cleaning/imputation/wrangling on previous version of the dataset.

Below is given name of file generated, the location of code in Python generating it, and its description about what it takes and what it does to it and hence generates the next version:

- **v0_deleted_extraneous_columns_and_date_format_fixed.csv [line # 24 – 45]**

  All columns other than the ones described previously have been omitted, and dates with format with hyphen and only 2 digits for year e.g. “03-05-18”, have been transformed into dates with format like “03/05/2018”. Also, ‘city’ column with HTML escape characters were fixed.

- **v1_inserted_countries_using_state-links.csv [line # 49 - 62]**

  There were several rows with missing values of ‘country’ , or ‘state/province’, or both, and of ‘UFO_shape’. Some of the ‘country’ entries were imputed with country codes, that were found with the same state-codes in other rows e.g. if in one row country and state/province were respectively ‘us’ and ‘ct’; and in another row it was ‘ ’ and ‘ct’, after imputation the later row had ‘us’ and ‘ct’ as country and state/province respectively. If same state-code was found with more than 1 countries, the imputation was ignored in that case.

- **v2_replaced_gb_with_uk.csv [line # 64 – 71]**

  The country code of UK, being “gb”, was replaced with “uk” because it aids in further extraction of country and/or state/province from the ‘city’ column.

- **v3_inserted_country_state_from_city_column.csv [line # 73 – 94]**

  Some ‘city’ entries had country and state/province in them in this format: “city(country/state-or-province)”. ‘country’ and ‘state/province’ columns were imputed with country codes and state/province names obtained this way.

- **v4_states_from_city_column_corrected.csv [line # 97 – 111]**

  “s.wales”, “wales, north country”, “n. ireland” were replaced with “south wales”, “north wales”, “northern ireland” respectively; all other entries having “wales” in state/province were imputed simply with “wales”; and corrupt forms of England like “englnd” and “endlnd” were fixed. Also, entries with UK areas like Birmingham, London were imputed with ‘uk’ as their country code.

- **v5_inserted_more_countries_from_city_column.csv [line # 114 – 130]**

  A list of country names and codes taken from the website countrycode.org, stored in ‘countrycode.org.csv’, was used. In the ‘city’ column, some entries held the country name in the format “city(country)”. The content within the last pair of brackets in each entry was checked if it had any country in the list acquired from the website, if so, then that row was imputed with the appropriate country code from the list that was acquired from the website.

- **v6_inserted_countries_from_city_column_using_whitespace_split.csv [line # 133 – 160]**

  The last pair of parenthesis in each ‘city’ entry was checked for having country name in format: “(<something><space>country<space><something>)” and if any match found in list of country names and codes obtained from countrycode.org, then the country column was imputed accordingly.

- **v7_removed_rows_with_missing_countries.csv [line # 163 – 170]**

  Any and all rows with missing ‘country’ field after all the previous steps, were omitted.

- **v8_set_uk_country_code_according_to_standards.csv [line # 173 – 177]**

  The country code of rows having country code of UK was set from “uk” to “gb”, since now the countries imputation is done and it needs to be replaced with the real country code for UK i.e. “gb”.

- **v9_state_column_fixed_for_uk.csv [line # 180 – 186]**

  All entries having “gb” as their country code, had their state/province fields values replaced like: “south wales”, “north wales” with “wales”; “northern ireland” with “ireland”.

- **v10_country_and_state_codes_replaced_with_names.csv [line # 189 – 209]**

  A list of country names, country codes, and their state codes and names, was obtained from this URL. The ‘state/province’ and ‘country’ fields were filled according to the country codes and state codes found in the list. The list of codes and names is also in the project subdirectory “loc172csv/2017-2 SubdivisionCodes.csv”

Using the Python Script (for desired visualizations):

The Python script takes the actual dataset file “ufo_sighting_data.csv” and performs all the previously explained steps for data cleaning, imputation, and wrangling, and generates corresponding versions. It also takes a file “visualization_params.csv” as input, telling it which city(ies), UFO_shape(s), state(s)/province(s), and country(ies) to include in the visualizations. A list of cities, states/provinces, UFO_shapes, and countries found uniquely in the dataset is in the project directory with name “uniques.csv”.

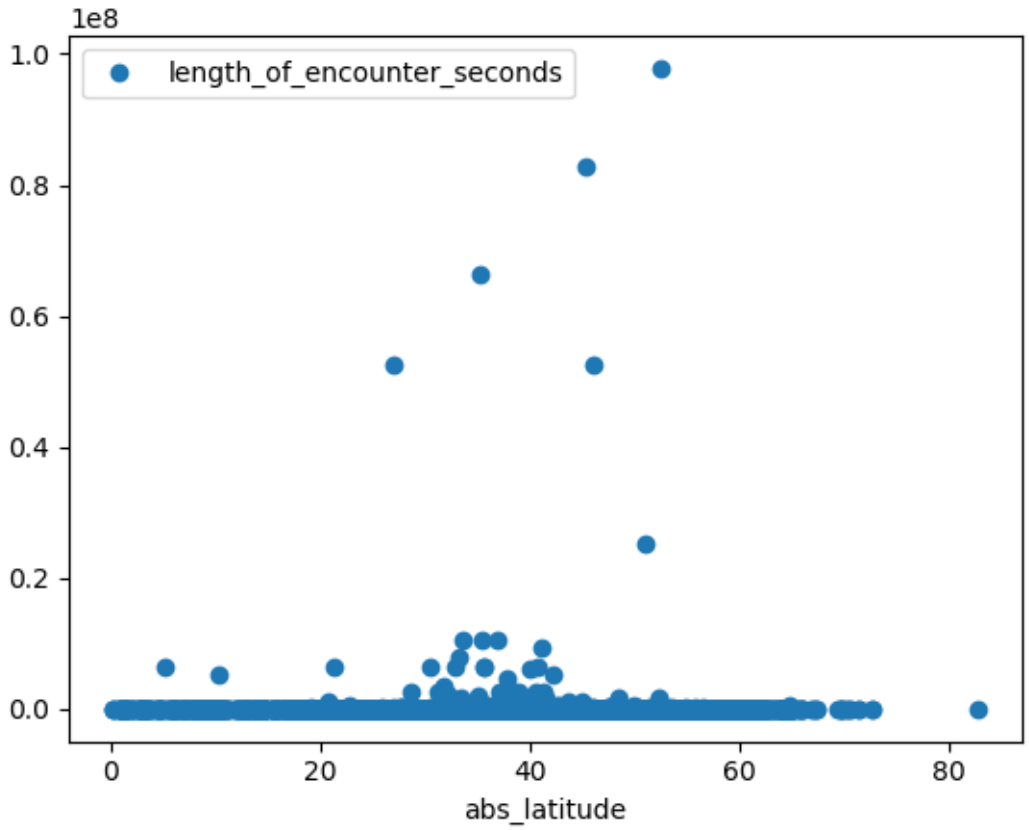
To specify which cities, states/provinces, countries, and UFO_shapes to include in the visualizations, simply look them up in the “uniques.csv” file, and fill them up in the appropriate columns of “visualization_params.csv”, and then run Python script “main.py”.

Some Sample Visualizations:

Some sample visualizations along with their insights are given below:

abs_latitude vs length_encounter:

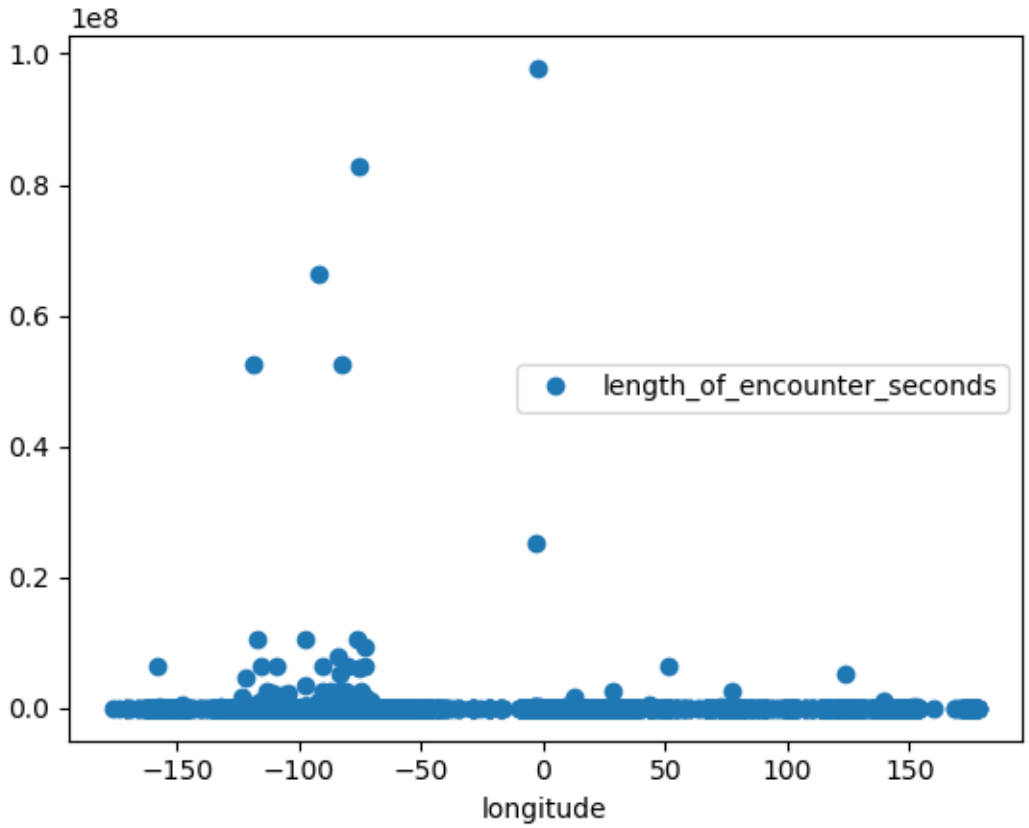
A scatter plot of absolute value of latitude coordinate of location, and length of UFO encounter in seconds, is given below:



The more close a location is to the equator, the more close its absolute latitude coordinate will be to 0, and the hotter the weather, or more intense exposure to sun radiations, will be of that location. So, the above plot shows that there is no correlation between the location's distance from the equator i.e. the aliens don't care if the weather is hot or cold, they'll come anyway :p. But there is some relatively greater accumulating around absolute value of 40.0, meaning that quite a few of them have been in moderately hot weathered areas.

longitude vs. length_encounter:

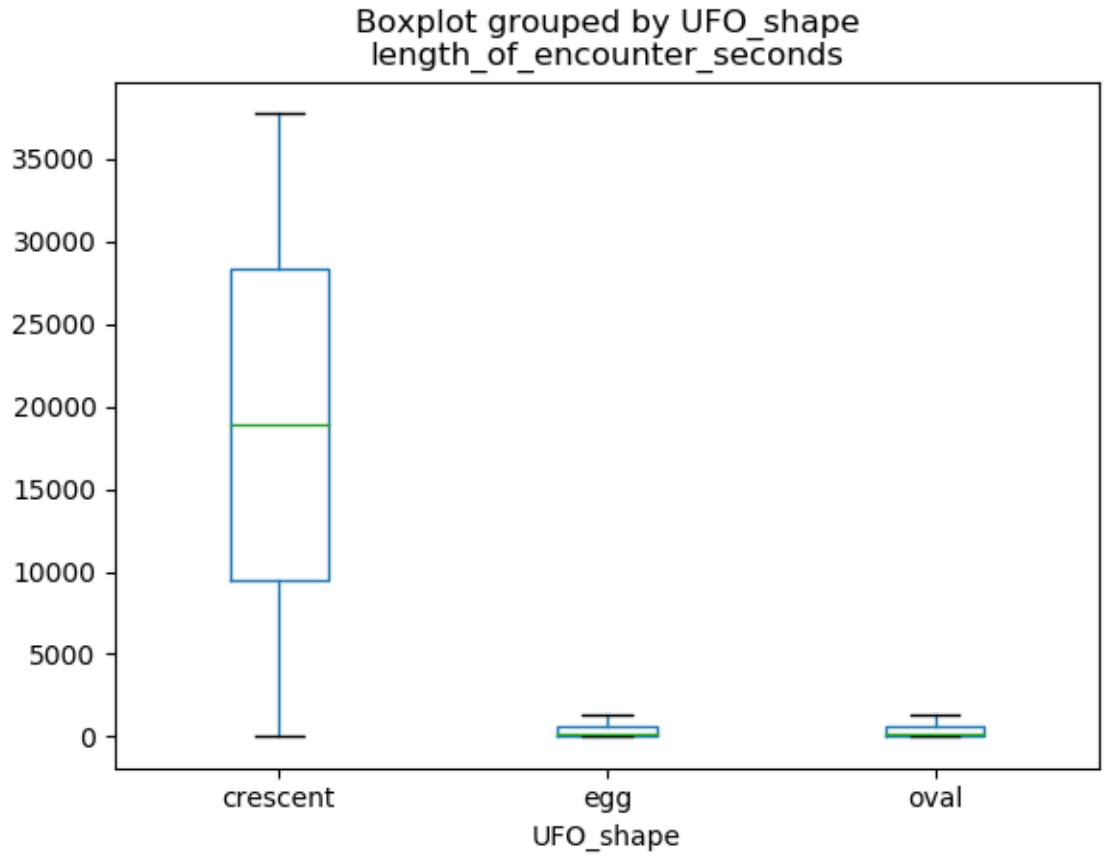
A scatter plot between longitude coordinate of location and length of UFO encounter in seconds is given below:



The more a location is to the east, the higher its longitude coordinate, and the more its close to the west, the lower it is. So, the slightly more high accumulation on the negative values indicate that more of the UFO sightings have been in the west. And that's no surprise, since aliens just prefer 'modern humans' :p.

UFO_shape vs length_encounter:

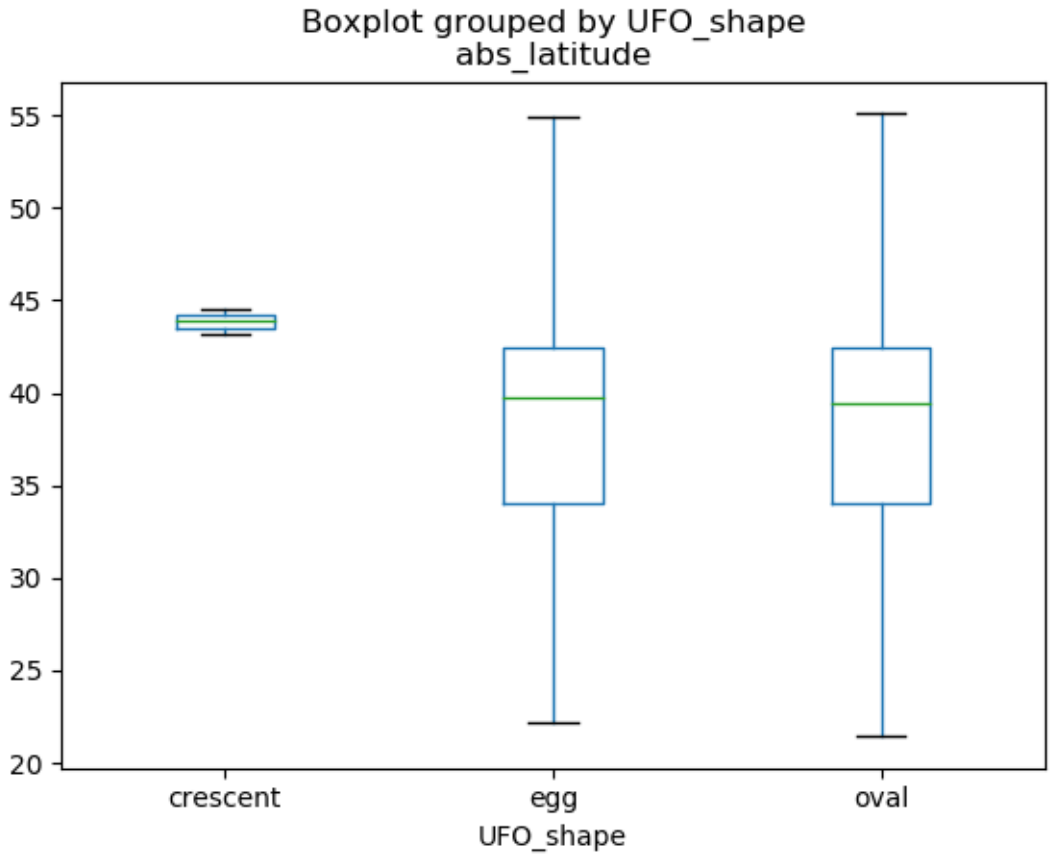
Below is given a set of box plots of different UFO_shapes, against their length of encounter in seconds:



This difference in sizes and shapes of the box plots tells us that sightings of some shape of UFOs were more lengthy, than others. And indeed, egg and oval shaped don't like staying near humans for long :p.

UFO_shape vs. abs_latitude:

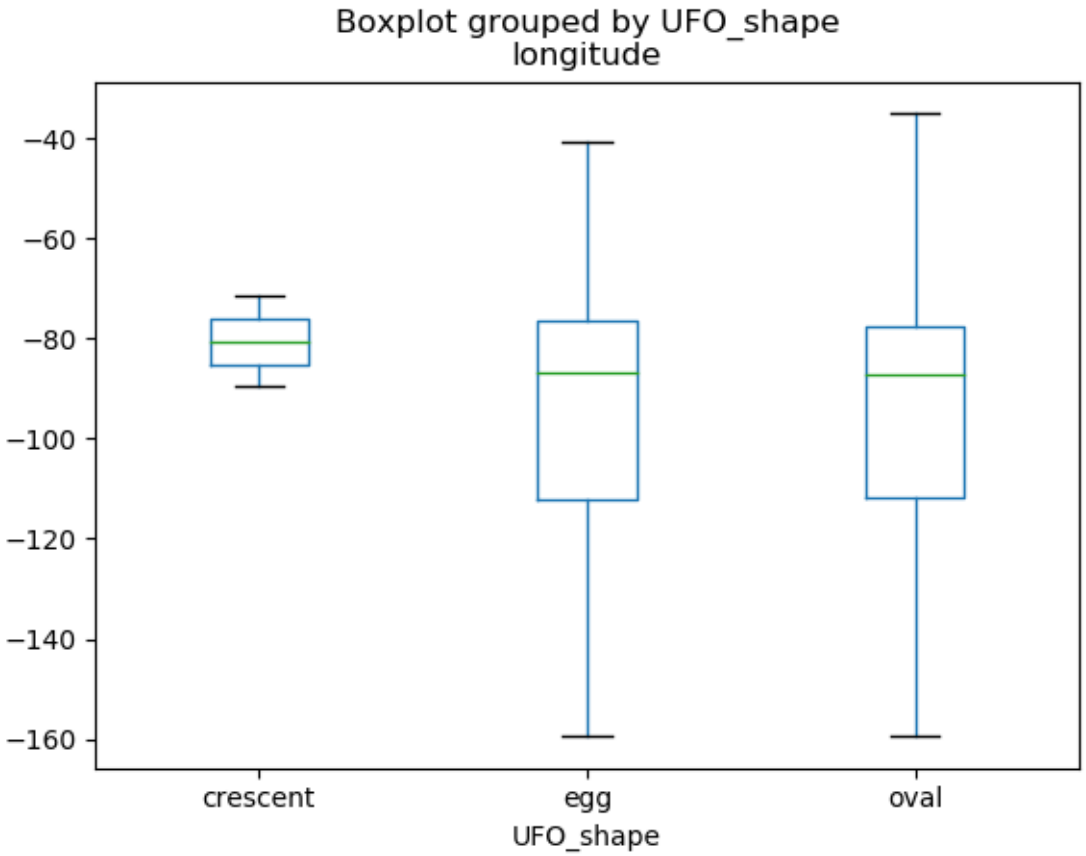
Below is given a set of box plots of different UFO_shapes, against their absolute latitude coordinate values:



This difference in sizes and shapes of the box plots tells us that sightings of what shape of UFOs have been more in areas exposed to more intense sun radiations or having hotter weather. From the look of things above, egg and oval shaped UFOs have quite a taste for a range of weather conditions, ranging from extreme hot to intense but not extreme cold; but I guess the crescent ones can't take the heat :p or cold ;)

UFO_shape vs. longitude:

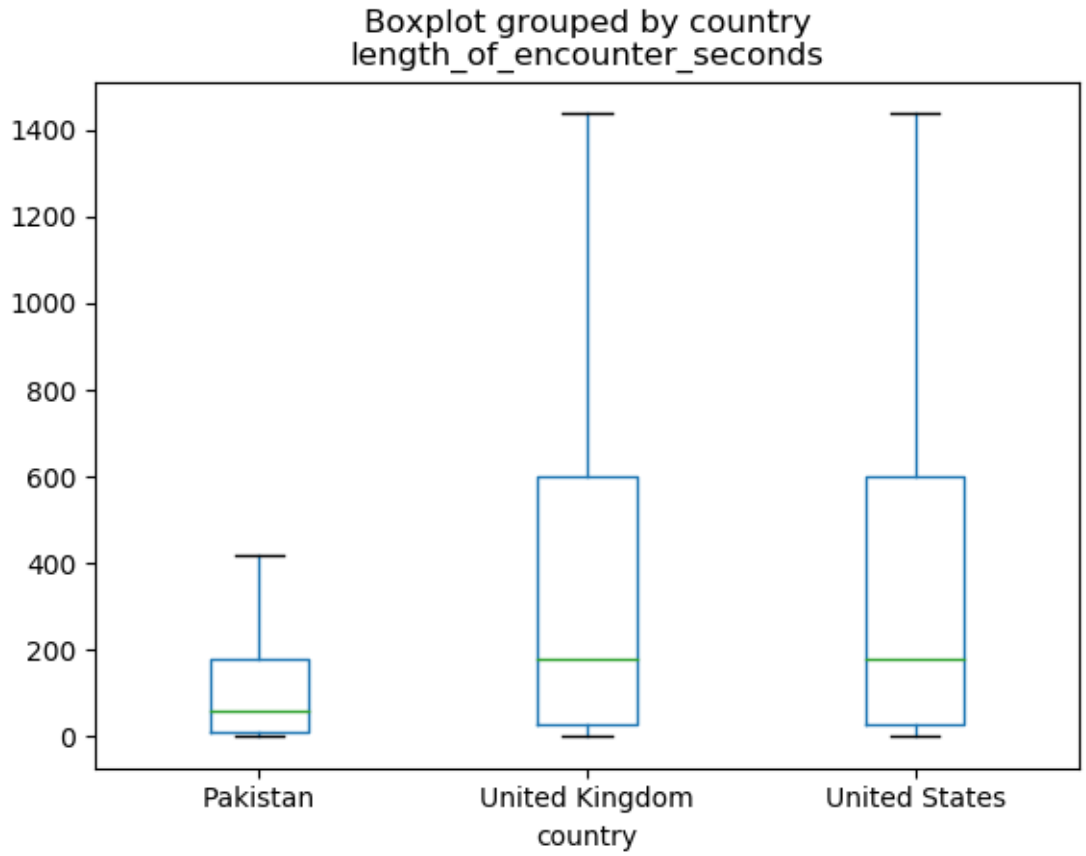
Below is given a set of box plots of different UFO_shapes, against their longitude coordinate values:



This difference in sizes and shapes of the box plots tells us that sightings of what shape of UFOs were more in which side of the world, east or west. From the plot above, clearly, aliens hangout a lot more in the west, and some like crescent ones, are even picky about where in west too :p.

country vs. length_encounter:

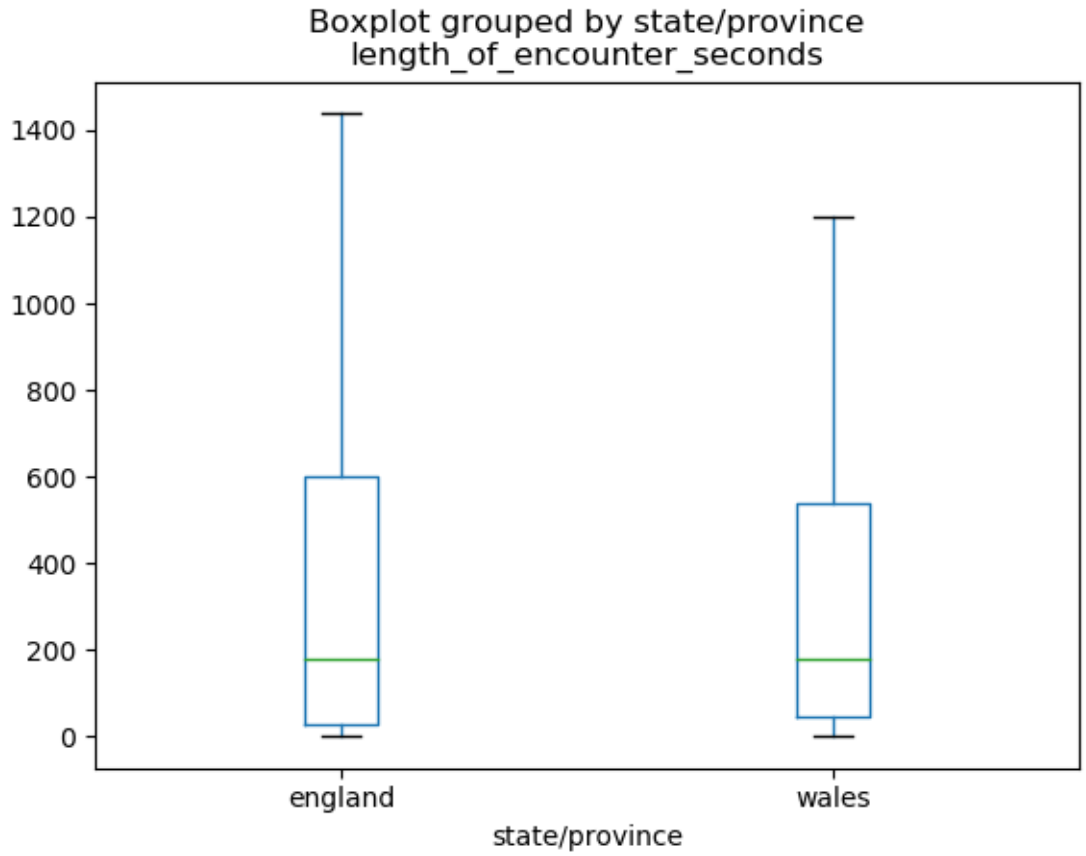
Below is given a set of box plots of different countries, against their length of UFO encounter durations in seconds:



This difference in sizes and shapes of the box plots tells us that some countries have had more lengthy sightings than others. Like above, our poor old homeland hasn't been visited much by UFOs or aliens :/. And when they did visit, the visits were quite short, as compared to UK and US :/.

state_province vs. length_encounter:

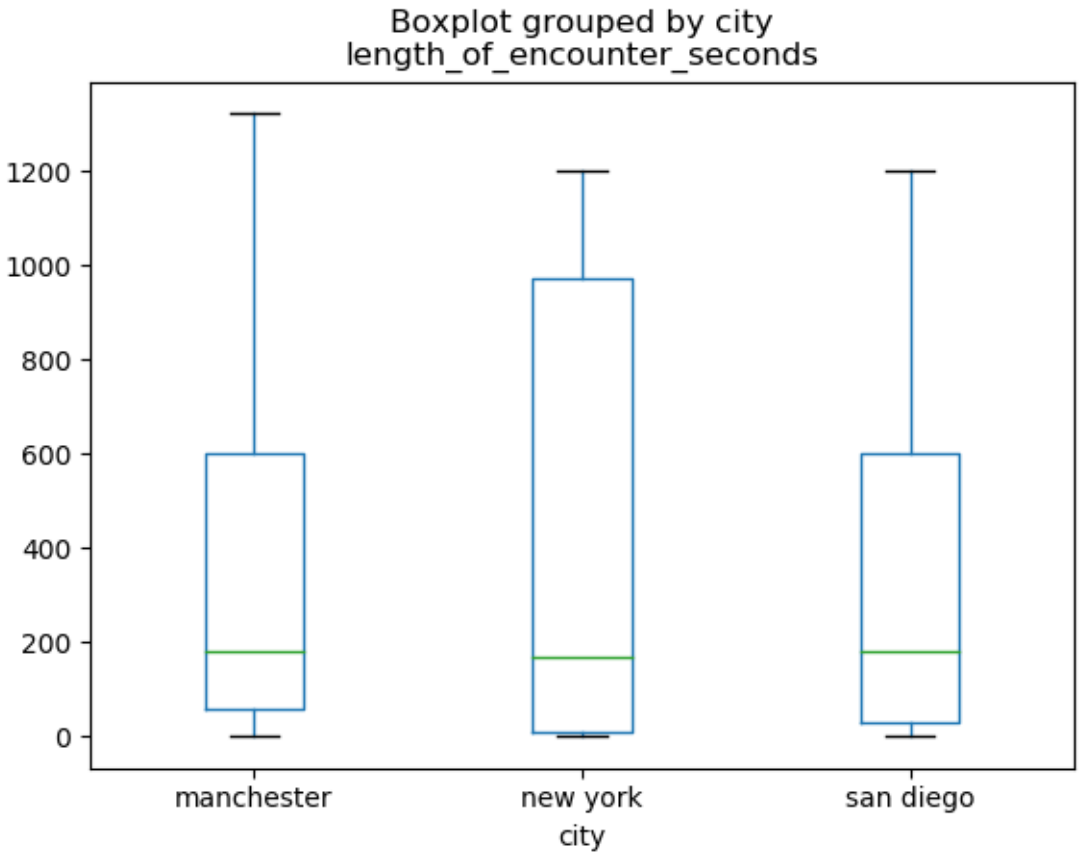
Below is given a set of box plots of different states/provinces, against their length of UFO encounter durations in seconds:



This difference in sizes and shapes of the box plots tells us that some states/provinces have had more lengthy sightings than others. Apparently, Wales isn't as comfortable as England for the aliens since they're maximum (of whisker) stay there is shorter :p

city vs. length_encounter:

Below is given a set of box plots of different cities, against their length of UFO encounter durations in seconds:



This difference in sizes and shapes of the box plots tells us that some cities have had more lengthy sightings than others. Like, Manchester has taken the lead in engaging the aliens for longest (according to whisker) of time interval :p