The Big Earth Data Project

Activities for Key Stage 5 (Using Python)

Activity set 1: Atmosphere – The Ozone layer

This set of activities explores the hole in the Ozone layer from 1980-2023 featuring daily and monthly data for the ozone hole and various atmospheric and weather measurements. The lessons cover aspects of the statistics element of A level Mathematics.

There is a very accessible guide to the hole in the ozone layer at https://discoveringantarctica.org.uk/oceans-atmosphere-landscape/atmosphere-weather-and-climate/the-ozone-hole/

All the lessons use Python *notebooks* (interactive coding documents featuring text and code). The notebooks can be accessed using Google Colab, a free, web-based platform. No prior knowledge of Python is assumed – all the code is given in the notebooks. Further support on using Python for data analysis can be found at: mei.org.uk/introduction-to-data-science/further-support/

We would value yours and your students' feedback. If time allows, please could your students fill in this feedback form (it does not have to be every student in your class) https://forms.office.com/e/UEGwYcanaz

The teacher feedback form is here, and also available on the website

https://forms.office.com/e/NGq9yr01pe

Overview of the lessons

Lesson 1: Understanding a context to solve a statistical problem/Interpreting averages and measures of variation

Lesson objectives:

- Explore the importance of understanding a context to solve a statistical problem.
- Learn how to use Python to explore a big data set.
- Interpret measures of central tendency and variation.

Activity overview:

- Introduction to the ozone layer and how satellite data is used in observing the hole.
- Getting started on using Python to explore data.
- Exploring the data sets to identify which periods during the year the hole in the ozone layer is observable.

Lesson 2: Diagrams for single variable data

Lesson objectives:

- Be able to interpret diagrams for single-variable data, including boxplots, histograms, frequency polygons and cumulative frequency curves.
- Consider the appropriateness of unfamiliar graphs or representations of data.
- Select or critique data presentation techniques in the context of a statistical problem.

Activity overview:

- Introduction to creating different diagrams for single variable data in Python.
- Using different diagrams to explore how the size of the ozone hole varies over different seasons and different decades.
- Considering which diagrams are the most and least useful.

Lesson 3: Exploring outliers

Lesson objectives:

- Recognise and interpret possible outliers in data sets and statistical diagrams.
- Determine whether an outlier is a valid data point.

Activity overview:

- Identify outliers in the monthly data for ozone hole size using the mean/standard deviation or quartiles/inter-quartile range.
- Consider whether an outlier is an error or valid data point.
- Find and classify outliers in the daily data.

Lesson 4: Identifying regions in scatter plots

Lesson objectives:

- Be able to interpret scatter diagrams for bivariate data.
- Be able to recognise distinct sections of a population within a scatter diagram.
- Use an informal understanding of correlation.

Activity overview:

- Plotting scatter diagrams and finding correlation coefficients in the monthly data to explore the factors associated with ozone hole size.
- Finding regions in scatter plots.
- Exploring the correlation in a region of a scatter plot
- Identifying regions in the scatter plots for the data since 1990

Lesson 1: Understanding a context to solve a statistical problem/Interpreting averages and measures of variation

Lesson objectives

- Explore the importance of understanding a context to solve a statistical problem.
- Learn how to use Python to explore a big data set.
- Interpret measures of central tendency and variation.

Notebook

Resource: https://colab.research.google.com/drive/1QZyphAr4UhfVhkEEBwDhEve8U8oAs11U

Lesson Plan

Exploring the importance of understanding a context to solve a statistical problem

- Introduce the importance of understanding a context when solving a statistical problem. This is essential to:
  - Clean the data. For example, a value of -1 for temperature in °C is appropriate but -1 for rainfall in mm isn't.
  - Choose appropriate statistics (such as averages) and charts to analyse the data.
  - Interpret the result: i.e. give a meaningful answer to a question.
- Introduction to the ozone context used in this set of lessons on 'Atmosphere'.
  - What is ozone?
  - What is the Ozone layer?
  - Why is it important?
  - Discovering the hole in the ozone layer
  - The development of the hole and the solutions in place.
- Introduction to the data sets
  - How the data was collected with satellites.
  - Information about the columns/variables in the data set. These will become more familiar when students have explored the data.
  - Show the change in the Ozone hole area over the year chart (from the appendix in the activity. Ask the students:
    - What do you notice about when the hole has been observable?

Learning how to use Python to explore a big data set.

- Importance of using technology to work with big data sets. Why use Python:
  - A lot quicker and can handle large data
  - An industry standard
  - Modelling/machine learning tools are built-in
- Introduction to using notebooks in either Kaggle or Google Colab:
  - Demo opening the notebook at https://colab.research.google.com/drive/1QZyphAr4UhfVhkEEBwDhEve8U8oAs11U

- Save your own copy:
  - Kaggle: *Copy and Edit*
  - Google Colab: *Copy to Drive*
- Always press 'Run all' before running any other cells.
- Demo changing a cell and running it: e.g. change the first code block from 1+1=2.
- Demo copying and pasting the code into a new block: copy the code to find the grouped statistics and change to 'month'.

- Students open the notebook and Copy, paste and edit the code to create grouped statistics and charts for months.
- Students use these statistics and charts to answer the **Task 1** questions (either individually or as a group discussion):
  - What season is the hole in the ozone layer mainly visible?
  - Why is the median a more useful average than the mean for summer, autumn and winter?
  - Why are September, October and November listed as 'spring?'
  - Which months show the biggest values and variation in the hole in the ozone layer? Is mean/std or median/iqr more useful for this?

Interpreting measures of central tendency and variation.

- Ozone depletion: Why does it change over the year?
- Line chart for stratosphere_temperature. This chart shows when the temperature is below -78°C. There is sufficient sunlight from August (1st August is day 213 or 214 of the year). (Code in the appendix of the notebook).
- Students complete **Task 2**
  - Students find the statistics and box plots for stratosphere_temperature by season.
  - Students find the statistics and box plots for stratosphere_temperature by month.
  - Questions:
    - How does the temperature in the stratosphere vary over the year?
    - Did you find the grouping by season or the grouping by month easier to interpret? There is no 'correct' answer to this – it is purely a preference!
    - Which average is more appropriate: mean, median or either?
    - Which measure of variation is more appropriate: standard deviation, inter-quartile range or either?
- **Extension task:**
  - Explore how some of the other numerical variables vary over the year.

Lesson 2: Diagrams for single variable data

Lesson objectives

- Be able to interpret diagrams for single-variable data, including boxplots, histograms, frequency polygons and cumulative frequency curves.
- Consider the appropriateness of unfamiliar graphs or representations of data.
- Select or critique data presentation techniques in the context of a statistical problem.

Notebook:

https://colab.research.google.com/drive/1smzwLk2-w5QiAyvd1mAKmtDOM4QnlItG

Lesson plan:

Creating diagrams in Python

- Introduce context – daily data for size of hole in Ozone area. NB the seasons are defined by southern hemisphere month, not solstices/equinoxes. Reminder of the time series for the ozone hole over the year.
- Visualisations: Charts and diagrams represent data graphically to emphasise the patterns in the data and make the patterns easier to understand: “a picture paints a thousand words”.
- In this lesson you will focus on how you can observe patterns in the data using the Seaborn visualisations package which can be imported into Python.
- Open notebook and ‘run all’
- Introduction to using Seaborn.
    - Define diagrams – note that these are all for a single data item
    - ‘Category’ plots
        - Box plot: min, lower quartile, median, upper quartile, max. Sometimes known as a 5-figure plot. Values more the 1.5 interquartile ranges above the upper quartile or below the lower quartile are shown as dots outside the whiskers.
        - Strip plot: A dot for each point. They are vertically spaced out to make it easier to read but the vertical position has no meaning.
        - Violin plot: a distribution curve estimating the *density* of the data at a value. Can be thought of as a histogram with infinitely thin bars.
        - Boxenplot: Similar to a box (the middle box shows the lower quartile to the upper quartile) outside of this the data is then split into half each time (the next two boxes show the 12.5-25 and 75-87.5 percentiles). The remaining data is then split in half each time.
    - Distribution plots:
      NB they can change binwidth in these.
        - Histogram: The standard histogram in Seaborn uses equal interval widths and frequency on the vertical axis. Note the use of binwidth and binrange.
        - Histograms can split or stacked into categories.
        - Frequency polygons use binwidth similarly.
        - Cumulative frequency is plotted as ‘ecdf’ and can have count or percent on the vertical axis.

Exploring the hole for different seasons

- Students complete **Task 1**:
  - They view the diagrams in the first section then answer the following questions:
    - How do the diagrams above show that the majority of the time that there is a hole in the ozone layer occurs in the southern hemisphere spring?
    - Which diagrams did you find the most useful? What features of the diagrams helped?
    - Which diagrams did you find the least useful?

Exploring the hole in the ozone layer for different decades

- Demonstrate filtering for Spring
  - Discuss how the Southern hemisphere Spring is where to focus your attention.
  - Explain the code that creates a new data set with just the Spring data.
- Students complete **Task 2**
  - They create some diagrams for the spring data grouped by decade.
  - Then answer the following questions:
    - How has the hole in the ozone layer changed over the last few decades?
    - Which diagrams did you find the most useful? What features of the diagrams helped?
    - Which diagrams did you find the least useful?
    - Why is the data for the 2020s different to the other decades? Which diagrams show this?

Extension task

- Explore the other two ozone columns for the different decades. Is it a similar pattern to the ozone hole area?
- Explore the minimum temperature for the different decades. Is there a pattern in this?

Lesson 3: Exploring outliers

Lesson Objectives

- Be able to locate outliers in single variable data.
- Consider whether an outlier is an error or a valid piece.
- Consider different approaches to dealing with outliers once identified.

Notebook

https://colab.research.google.com/drive/1Wz4nehnArSDKMyifIfxBuI4WZtywf5Lu

Lesson plan

Identifying an outlier

- Remind of the context of the ozone hole, particularly its seasonality.
- Show the monthly data. Note that for the monthly data the measurements in the right-hand column are the means of the daily values.
- Open notebook and run all. Show the box plot for the ozone hole area and discuss the outlier for December. NB The convention used here is to mark points more than 1.5 inter-quartile ranges above the upper quartile as points outside the 'whiskers'.
- Definitions of outliers. Conventionally the quartile rule is nearly always 1.5 IQR. For the mean and standard deviation different amounts of the sd can be used. These are 'rules of thumb' and can vary due to the context. Students will explore different ones in this lesson. Show that the ozone hole area value for December 2010 is an outlier using either of the methods. NB Python can be used as a calculator.
- Students discuss the questions in **Task 1**:
    - Explain why it will not be possible for there to be a 'lower' outlier for the December ozone data using these boundaries.
    - Is the outlier displayed in the box plot in December an outlier for both definitions used?
    - Other definitions can also be used. Will the outlier displayed in the box plot in December still be considered an outlier if you use 3 standard deviations above the mean as a boundary condition?
    - Why do you think this outlier has occurred? What should be done about it?

Classifying an outlier: error or valid data point

- Discuss one of the reasons for finding outliers or valid data points is to identify if it is an error or a valid data point.
- You can do this by exploring the other variables in this context to help you decide if an unusual value for December 2020 is reasonable. Show the slice for recent years (since 2015) and how to create time series for other variables.
- Students complete **Task 2**:
    - Find some time series for the other variables for recent years
    - Answer questions about what they've found
        - Was there anything in any of the other measures that made 2020 unusual?

- Do you think the value for ozone_hole_area in December is an error or a valid data point?
- How have the other variables helped you make your decision?

Finding a classifying other outliers in the monthly data

- One of the other months that had a noticeable outlier was May. Demonstrate filtering the data for May.
- Students complete **Task 3**:
    - Find the year that has an outlier for May.
    - Filter the data to find the decade that has this outlier and explore the other variables for this month.
    - Answer questions.
        - What was different about the ozone hole your chosen year?
        - Was it particularly different in any other categories?
        - Do you think the data is valid or an error?

Finding and classifying errors in the daily data

- Create a set of box plots for the daily data for different features.
- Identify the large outliers in temperature and ozone column minimum.
- Answer questions.
    - Which features and dates of the daily data had large outliers?
    - Why do you think these outliers have occurred?
    - What should be done about these data points?

Lesson 4: Identifying regions in scatter diagrams

Lesson objectives

- Be able to interpret scatter diagrams for bivariate data.
- Be able to recognise distinct sections of a population within a scatter diagram.
- Use an informal understanding of correlation.

Notebook:

https://colab.research.google.com/drive/1obL6zB39W92NjqFaeThGzMlk2OwfYcU3

Lesson plan:

Plotting scatter diagrams and finding correlation coefficients

- Introduce context – monthly data for size of hole in Ozone area. In this activity students will explore whether there is a link between any of the weather or atmospheric measurements and the size of the hole in the ozone layer.
- Open notebook and 'run all'
- Discuss how the boxplots show the season has a much bigger impact on the hole in the ozone layer than the decades. The code for filtering for spring is shown.
- Demonstrate how to find a scatter plot for ozone hole versus air temperature and how to find the correlation coefficient.
- Students complete **Task 1**:
    - Find scatter plots and correlation coefficients for ozone hole vs the other three atmospheric readings (surface_pressure, wind_speed, stratosphere_temperature).
    - Discuss the following questions:
        - What type of correlation is shown in each diagram (positive/negative)?
        - Which show that strongest correlation?

Finding regions is scatter plots

- Demonstrate how to add a third variable to a scatter diagram. Note that for month you need to add in a 'palette' to make the colours distinct instead of a scale. If students have difficulty distinguishing between the colours see the alternative approaches in the appendix of the notebook.
- Students complete **Task 2**:
    - Create some scatter plots and calculate correlation coefficients to compare ozone_hole_area to the other atmospheric and weather condition, grouped by either decade or month.
    - What patterns are there in the data?

Exploring the correlation in a region of a scatter plot

    - Demonstrate filtering the data for September and plotting a scatter diagram/finding a correlation coefficient.
    - Students complete **Task 3**:

- Find the scatter plots and correlation coefficients for ozone_hole_area versus the other weather/atmospheric variables for September. Which shows the strongest correlation? How does this compare to the correlation for the data for the whole of spring?
- Repeat this for the October and November data.

Extension: Identifying regions in the scatter plots for the data since 1990

- Demonstrate finding the scatter plot and box plots grouped by decade.
- Discuss how the scatter plot and box plot relate to each other. Why does this suggest that it might be useful to concentrate on the data from 1990 onwards?
- Demonstrate filtering the data for 1990 onwards and finding some scatter diagrams/correlation coefficients.
- Students do the **extension task**:
  - Explore the strength of the correlation for ozone_hole_area versus the weather/atmospheric variables if only the data from 1990 onwards is used.