**Lab 4 – Data Analysis**

**Part 1 – Understanding Graphs**

Ref: http://www.statisticalsolutions.ie/wp-content/APPLIED%20STATISTICS%20ebook.pdf

**Chart, histogram

Description automatically generated**

**Chart, histogram

Description automatically generated**

**Chart, line chart

Description automatically generated**

Review the last 30 days, finishing yesterday, and record the daily values of a variable e.g. the number of messages received that day on a messaging app and then do the following:

1. Define the variable of interest
2. Show the data
3. Draw the time series plot (by hand and then at the end of the lab in python!)
4. Identify any notable features in the shape of the plot
5. Explain in simple language what these features are telling you about this variable.

**Part 2 – Plotting Graphs in Python**

*This Lab is an extension of Lab 3 – so the parts in BLUE are the new elements. However if you did not get all questions answered for this last week you should do it now!*

**Question 1 – Read in and display a test .csv file**

Go to moodle and download the weather.csv file.

Have a look at the data in this file. You are to write a paragraph explaining this data. What is each element and why you think it is that, give your reasoning.

Write and test the python code to do the following:

* Read the .csv file into python
* Put the .csv file into a suitable list
* Print out that list to the screen
* Chart the Temperature versus the humidity
* Convert the Temperature to Fahrenheit and re graph the two sets of data.
* Write a few sentences on what both graphs mean.

**Question 2 – Google Play Store**

Go to moodle and download the googleplaystore.csv.

**Background:**

The details of the applications on the google play store. The data is from a specific date and is only a sub-section of all the data that was allowed: The columns are as follows:

**App:** Application name

**Category:** Category the app belongs to

**Rating:** Overall user rating of the app (as when scraped)

**Reviews:** Number of user reviews for the app (as when scraped)

**Size:** Size of the app (as when scraped)

**Installs:** Number of user downloads/installs for the app (as when scraped)

**Type:** Paid or Free

**Price:** Price of the app (as when scraped)

**Content Rating:** Age group the app is targeted at - Children / Mature 21+ / Adult

Genres. An app can belong to multiple genres (apart from its main category). For eg, a musical family game will belong to Music, Game, Family genres.

**Last Updated:** Date when the app was last updated on Play Store (as when scraped)

**Current Ver:** Current version of the app available on Play Store (as when scraped)

**Android Ver:** Min required Android version (as when scraped)

Write and test the python code to do the following:

1. Read the .csv file into python
2. Put the .csv file into a suitable list
3. Print out that list to the screen only print the first 20 rows and 8 attributes as this is a very big data set.
4. Display the total free games downloaded versus the total paid games downloaded.

4a. Show the above results in a bar chart.

1. Display the percentage of free games downloaded versus the percentage of paid games.

5a. Show the above results in a bar chart. What do you notice about this chart and the previous one? Explain why this happens.

1. Calculate and display the average price of an app.
2. Calculate and display the minimum and maximum number of installs for an app – you don’t need the app name but if you want an extra challenge you can include it!

7a. Graph the Average, Max and Min number of installs from above.

1. Calculate and display the minimum and maximum ratings for an app – you don’t need the app name but if you want an extra challenge you can include it!
2. Display the number of installs of free apps (this can be a list I don’t need a total here) and the number of installs of paid apps (this can be a list I don’t need a total here). For an extra challenge get the totals i.e. total number of free app downloads.

9a. Graph the data that shows all the free downloads and all the paid downloads. So the graph should have an entry for each game and it should display downloads for this game. On the chart we should see two lines one to represent the free games and one to represent the paid games. Explain what we see on this chart.

1. Display the average number of free apps and the average number of paid app installs.

10a. Show the above results in a bar chart.

If you have completed all of the above go back and take on the extra challenges if you did not try them already.

**Question 3 – Wine Ratings**

The data in this data set has the following columns

**country**

The country that the wine is from

**description**

designation

The vineyard within the winery where the grapes that made the wine are from

**points**

The number of points WineEnthusiast rated the wine on a scale of 1-100 (though they say they only post reviews for wines that score >=80)

**price**

The cost for a bottle of the wine

**province**

The province or state that the wine is from

**region\_1**

The wine growing area in a province or state (ie Napa)

**region\_2**

Sometimes there are more specific regions specified within a wine growing area (ie Rutherford inside the Napa Valley), but this value can sometimes be blank

**taster\_name**

taster\_twitter\_handle

**title**

The title of the wine review, which often contains the vintage if you're interested in extracting that feature

**variety**

The type of grapes used to make the wine (ie Pinot Noir)

**winery**

The winery that made the wine

Write and test the python code to do the following:

1. Read the .csv file into python
2. Put the .csv file into a suitable list
3. Print out that list to the screen only print the first 20 rows and 8 attributes as this is a very big data set.
4. Display the prices and the points for all the wines in the dataset

4a. Graph both of these on a line graph. Write a paragraph about the chart and include this in your solution. Write an explanation for what you are seeing in the chart.

1. Display the Average and median prices and points and also display the mode points.
2. Show the prices and points for all wines from Northern Spain.

6a. Graph both of these on a line graph. Write a paragraph about the chart and include this in your solution. Also compare this chart to the previous one.

1. Calculate and display the Average and median prices and points and also display the mode points from Northern Spain.

7a. Show the above data on a Bar Chart

Question 4

Get the “Primary Out of School” excel file down from moodle.

What data does it contain?

Create a .csv file of this data.

Using Python do the following:

* Read this data into a suitable list
* Using a ‘Line graph’ graph the list.
* Get the average and median values and display these on a bar chart.
* Go to google public data and down load the full set of data for children out of school.
* Open the excel file and get the data for children in the United Kingdom.
* Add this data to your .csv file.
* Re-read in this data.
* Graph, using line charts the two sets of data.
* Write a paragraph on the differences, or not between these sets of data.
* Now go back to the full set of data and get the data for children in Latin America.
* Add this data to your.csv file and draw the graph to include this new data.
* What conclusions can we draw from this final chart? Spend a good paragraph describing this.