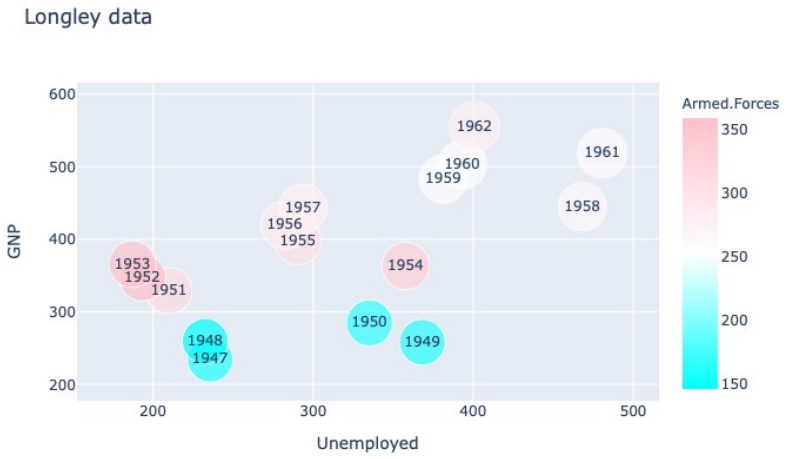
Exercise 1 Let's visualize Longley's macroeconomic data set. Google longley-dataset to read more about this data. You can either read the excel file on moodle or install the package `pydataset` and access the data like this

```
from pydataset import data
longley = data('longley')
```

Plot the number of unemployed persons (variable `Unemployed`) against the gross national product (variable `GNP`), label each point with its year (variable `Year`), the size of each point should be the population size (variable `Population`), and the color corresponds to the number of people in the armed forces (variable `Armed.Forces`). Your result need not be exactly the same but should be similar to the following plot.



Exercise 2 Statistics needs data. Unfortunately, data often cannot be collected fully. Therefore many data sets contain „gaps", non-existing measurements, so-called `NA`'s (not available). We will consider here an example.

The data set `iris` contains measurements of the length and the width (in cm) of petals and sepals of three iris species: 1: Setosa, 2: Versicolor and 3: Virginica. This data set `iris` is already part of the standard `plotly.express` installation. Before you start, make a copy of the data set by `df = px.data.iris()`. Cast the data to `DplyFrame(df)` and look at the first few rows.

(i) How is this data set structured? How many observations (rows) does it contain? How many variables (columns)?
Hint: Use `shape` from `pandas`

(ii) To get an overview of the range of values, use the `pandas` function `describe()` on the data set. What information do you again?

(iii) Assume that we were unable to take the second observation of `petal_length` and `petal_width`, and for the fifth observation, the data for `sepal_length`, `sepal_width` and `petal_width` are missing. Replace these five fields by `None`.

Hint: Use the following command for the first two replacements. Figure out how to do the other three replacements.

```
df.loc[[1], ['petal\_length', 'petal\_width']] = None
```

(iv) Is there a difference is you again use `describe()` on the data set?

(v) Why should missing values always be coded by `None`, and not, for instance, filled with a zero? Explain it for the case of the `mean()` function.

(vi) The function `dropna()` eliminates all observations from the data frame for which any(!) variable contains `NA`'s. Save the result of `dropna(df)` in a new `data.frame`. How many observations remain?

(vii) Interpolate the missing values using the function `df.interpolate()`. Explain what you observe. Is it a good idea to deal with missing values in this particular case? Explain your reasoning.

Exercise 3 We will again look at the `iris` data set.

(i) Reload the data set to be sure that you use the original data.

(ii) Draw a scatter plot of the variables `sepal_width` and `sepal_length`. Do you see a linear relationship? Describe your observation.

(iii) Draw a scatter plot of the variables `petal_width` and `petal_length`. Do you see a linear relationship? Describe your observation.

(iv) Make the same two plots as before but this time color the points with respect to the corresponding species. What are your observations?
Hint: Use `color= ‘‘species’’`

(v) To display all variable combinations make a scatter plot matrix with `scatter_matrix`.

Exercise 4 The Effect of Vitamin C on Tooth Growth in Guinea Pigs is the topic of the `ToothGrowth` data set. Google the dataset to read more about it. You can either read the excel file on moodle or install the package `pydataset` and access the data like this

```
from pydataset import data
tooth = data('ToothGrowth')
```

We would like to see if the variable `supp` as well as the `dose` has an effect on tooth growth in Guinea Piges. Create a boxplot of the data and try to figure out how to present all the information in one plot in a neat way. Interpretate your plot.

Exercise 5 Load the `USArrests` data set, also part of the the package `pydataset` or available on moodle. Read about the data set on Google.

(i) Make a boxplot of the data (use `px.box`) for the variables `Murder`, `Assault`, and `Rape`.

Hint: you need to change the shape of the data.

(ii) How many counties have more than 10 Murders cases, more than 300 Assaults cases ***or*** more than 10 Rape convictions per 100'000 cases?

(iii) Which counties have more than 10 Murders cases, more than 300 Assaults cases ***and*** more than 10 Rape convictions per 100'000 cases?