# OLA 1

Christoffer Drejer Mikkelsen

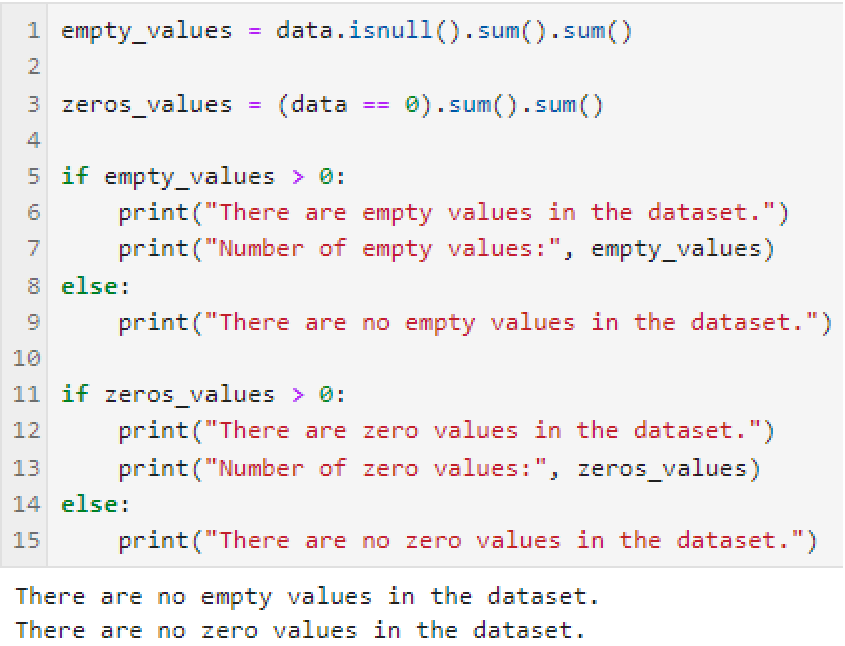
Cph-cm370@cphbusiness.dk

## Task 1: Data Exploration and Cleaning

The dataset is salary data from jobs within Data Science from the years 2020 – 2023.

https://www.kaggle.com/datasets/hummaamqaasim/jobs-in-data

No null values or zeros found in the dataset.



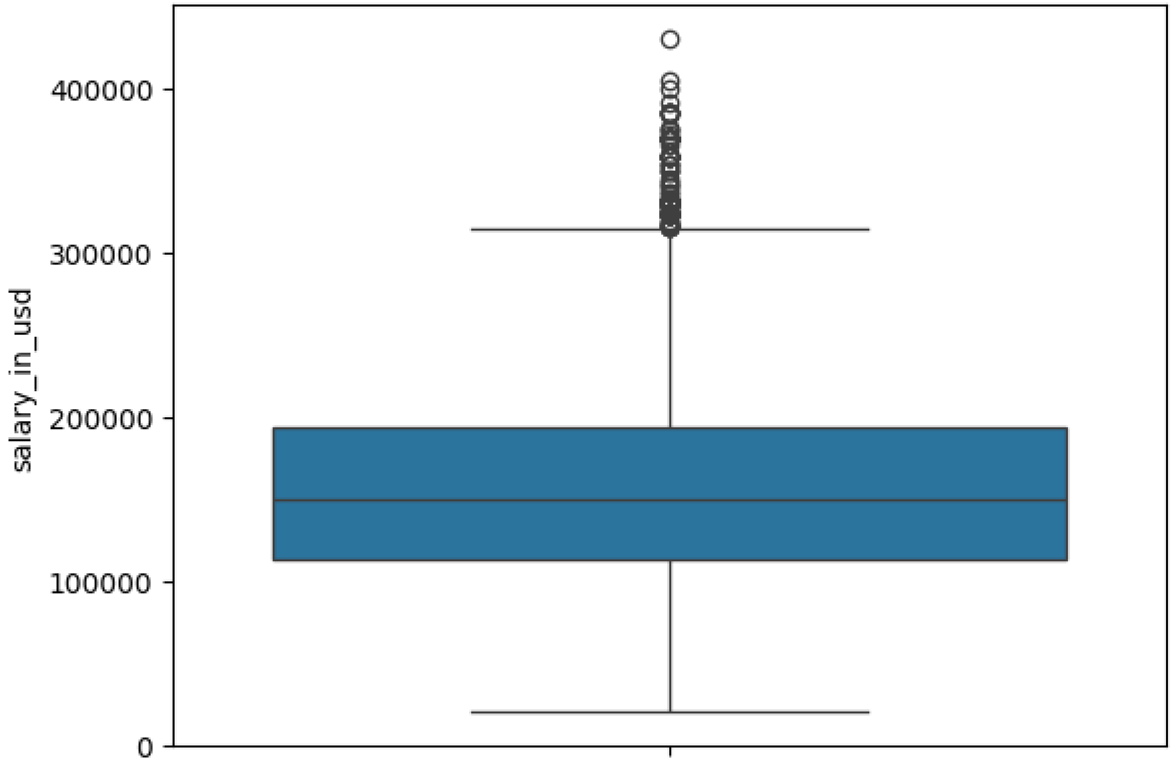
```
empty_values = data.isnull().sum().sum()

zeros_values = (data == 0).sum().sum()

if empty_values > 0:
    print("There are empty values in the dataset.")
    print("Number of empty values:", empty_values)
else:
    print("There are no empty values in the dataset.")

if zeros_values > 0:
    print("There are zero values in the dataset.")
    print("Number of zero values:", zeros_values)
else:
    print("There are no zero values in the dataset.")
```

```
There are no empty values in the dataset.
There are no zero values in the dataset.
```

Majority of the data is from 2023 so therefore it might be interesting to only look at the data from that year.

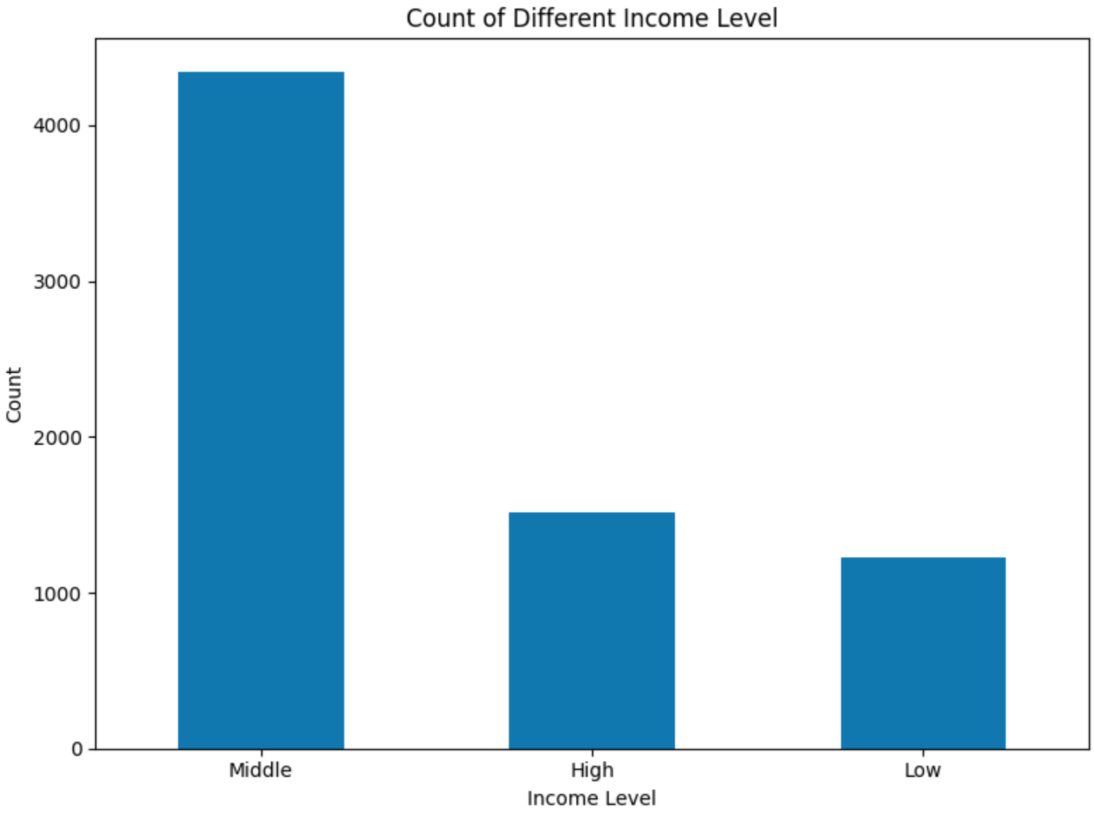
Most of the data is from people working in the US but we include the highest 5 countries so that we can make comparisons.

Using seaborn boxplot we find that we have some outliers when it comes to salaries.

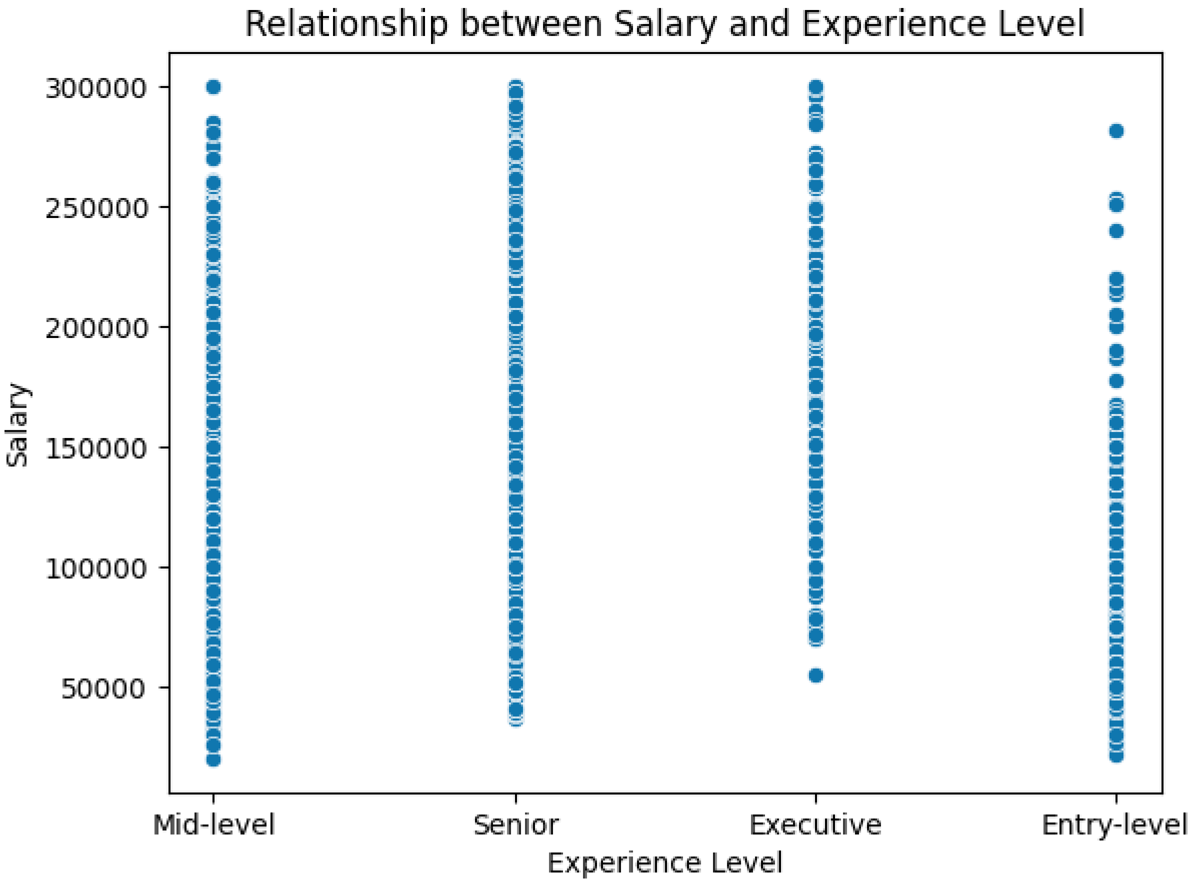


The outliers are on the higher end and earning above 300.000 USD per year and we therefore exclude this data.

We create a new feature that put salaries into three different categories.



We use seaborn scatterplots to find relationships between different data such as salary and experience.


Relationship between Salary and Experience Level
Salary
Mid-level
Senior
Executive
Entry-level
Experience Level