**Step 1: Introduction**

Title: Exploring Data Analysis and Visualization in Python

This dataset appears to represent a sample of individuals along with various attributes related to their demographics, education, work experience, and location. Here's a brief introduction to the data:

* Gender: Indicates the gender of each individual, with options being male or female.
* Salary: Represents the salary earned by each individual.
* Education Level: Describes the highest level of education attained by each individual, ranging from high school to PhD.
* Years of Experience: Indicates the number of years of professional experience each individual has.
* City: Specifies the city in which each individual resides or works.
* Selected: Indicates whether each individual has been selected for some criteria or not

**Slide 2: Loading and Inspecting Data**

* Code snippet importing necessary libraries: pandas, numpy, seaborn, matplotlib, plotly, etc.
* Reading data from a CSV file (Selection\_Data.csv) using pd.read\_csv()
* Displaying basic information about the dataset using df.info()

**Slide 3: Data Cleaning**

* Handling missing values: Checking for NaN values using df.isna().sum() and dropping them using df.dropna(inplace=True)
* Filtering out negative values in the 'years\_of\_experience' column
* Checking and removing duplicate rows using df.duplicated().sum()

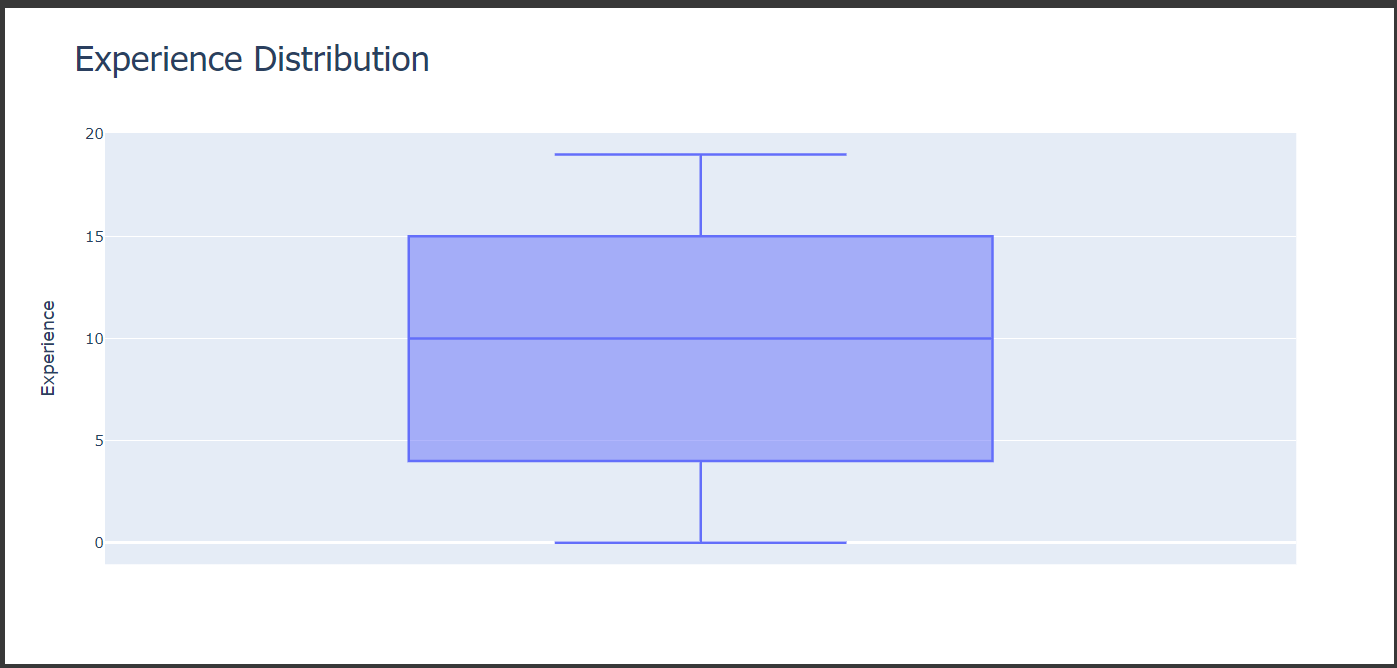
**Note: have written df = df[df['Selected'] == True] since I did it without filtering on 'Selected'. However, the visualizations were not accurate, and all the values, for example, salary versus education level, showed the same level bars in a bar graph when 'Selected' was not filtered. Hence, we have filtered it.**

**Slide 4: Data Preprocessing**

* Converting 'birth\_date' column to datetime format
* Calculating age of individuals based on birth date and current date
* Resetting DataFrame index after data manipulation

**Slide 5: Exploratory Data Analysis (EDA)**

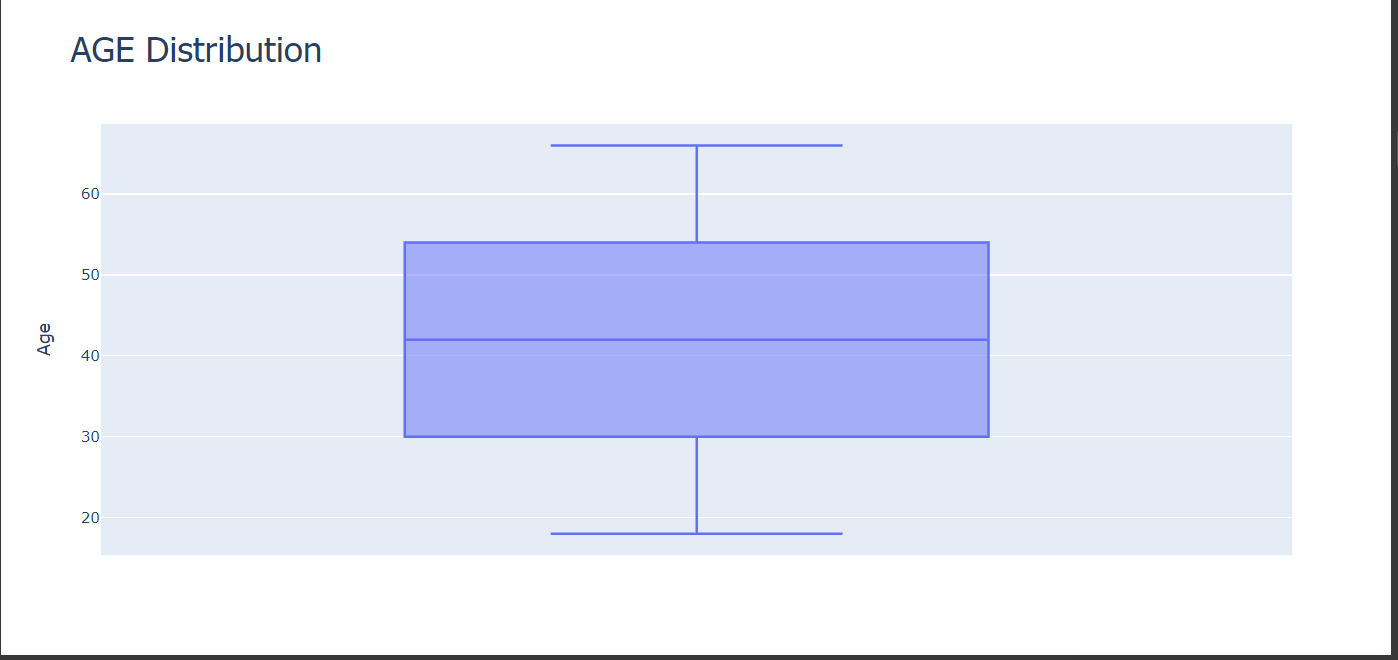
* Visualizing 'years\_of\_experience' distribution using a box plot with Plotly
* Customizing layout for better visualization experience
* Displaying the box plot using iplot(fig)



**Conclusion**: From the above diagram, we can conclude that the lowest experience is 0 years, the highest is 19 years, and the median is 10 years.

**Slide 6: Further EDA**

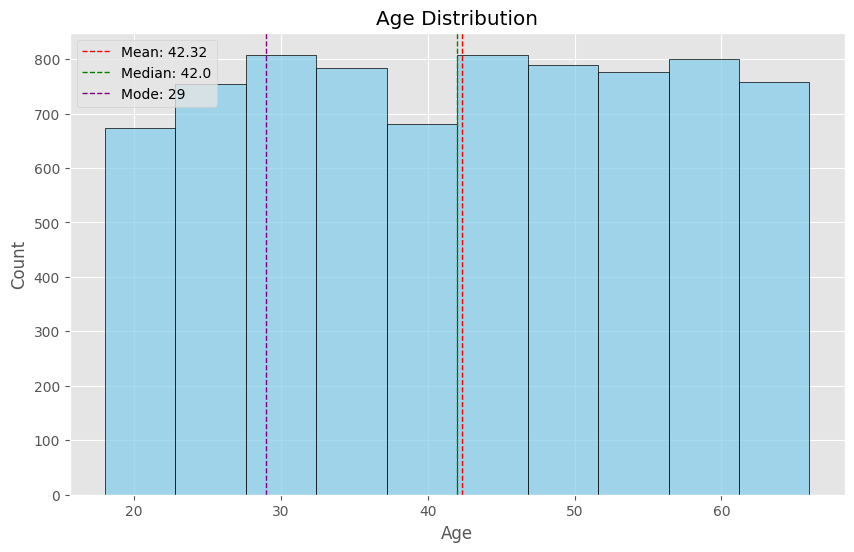
* Visualizing 'age' distribution using another box plot with Plotly
* Customizing layout for a dark theme
* Displaying the box plot using iplot(fig)

****

Conclusion: From the above diagram, we can conclude that the lowest age is 18 years, the highest is 66 years, and the median is 42 years.

**Slide 7: Additional Visualization**

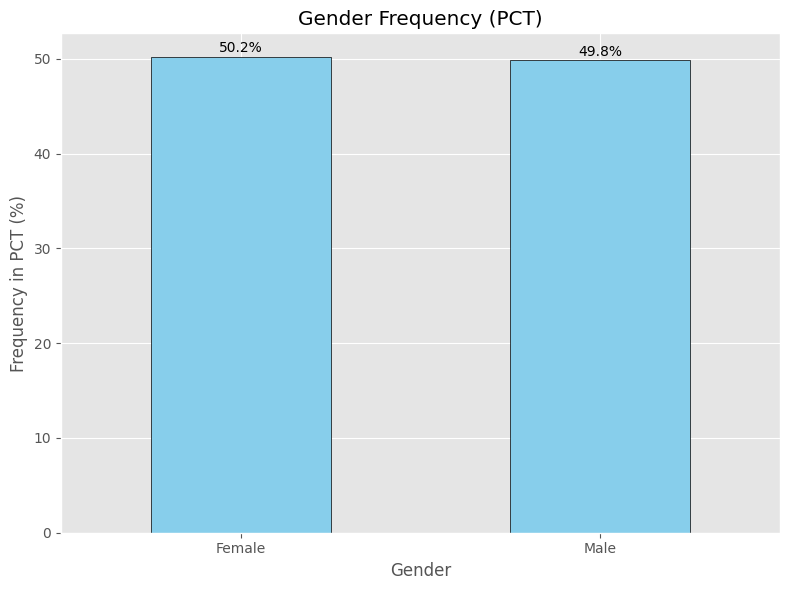
* Creating a histogram to visualize 'age' distribution using Seaborn
* Adding mean, median, and mode lines for better interpretation
* Displaying the histogram using plt.show()



**Conclusion: We get the mean median and mode from the ages**

**Slide 8: Analyzing Categorical Variables**

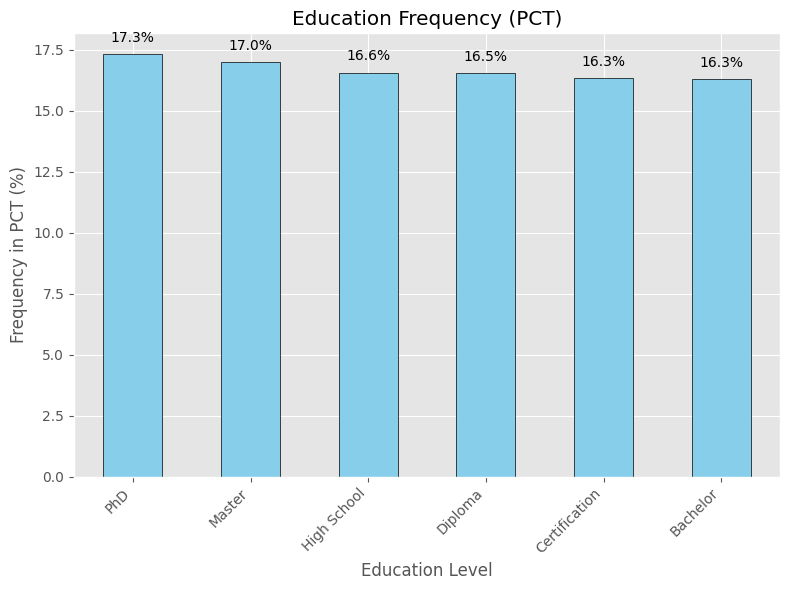
* Calculating and visualizing frequency distribution of 'gender' using a bar plot
* Converting frequency counts to percentages for better comparison
* Displaying the bar plot using plt.show()

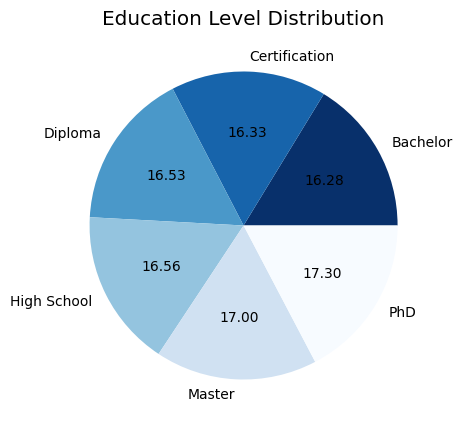


Conclusion: We can also see that we have an even Number of Male & Females. (A Very slight difference but it does not look skewed)

**Slide 9: More Analysis on Categorical Variables**

* Calculating and visualizing frequency distribution of 'education\_level' using a bar plot
* Similar process as previous slide but for education level
* Displaying the bar plot using plt.show()

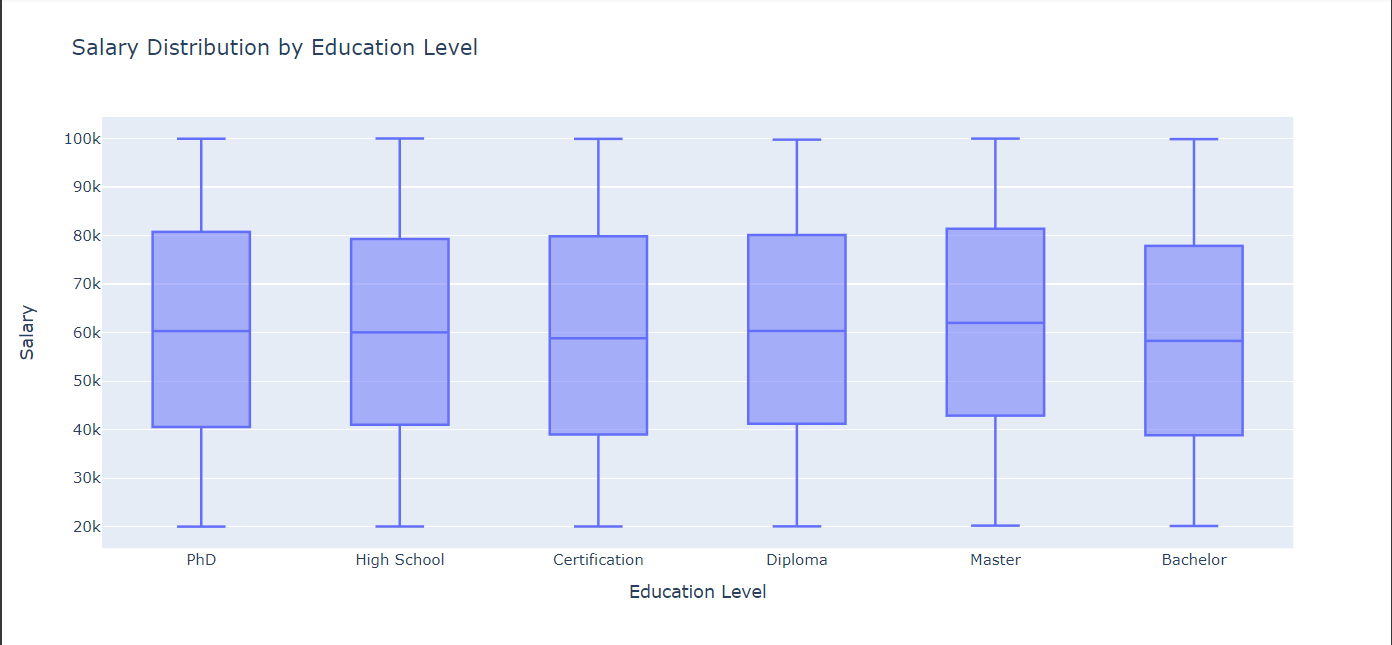




**Conclusion:** We have see that we have PhD>Masters>High School> Diploma>Certification> Bachelors

**Slide 10: Relationship Analysis**

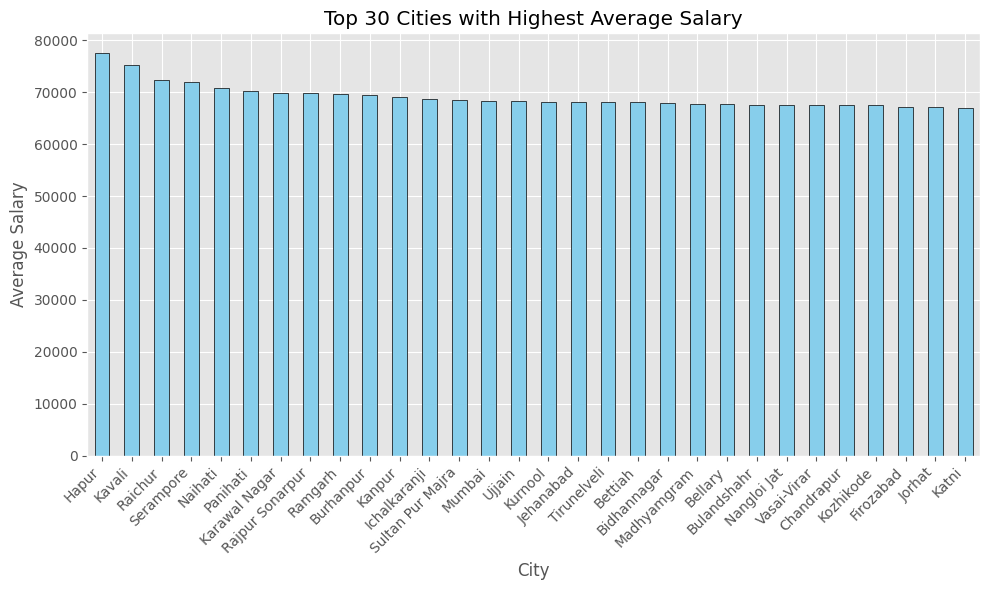
* Visualizing the relationship between 'education\_level' and 'salary' using a box plot
* Customizing layout for better presentation
* Displaying the box plot using fig.show()



**Conclusion:** We have see that we have Masters has the highest pay

**Slide 11: Geographic Analysis**

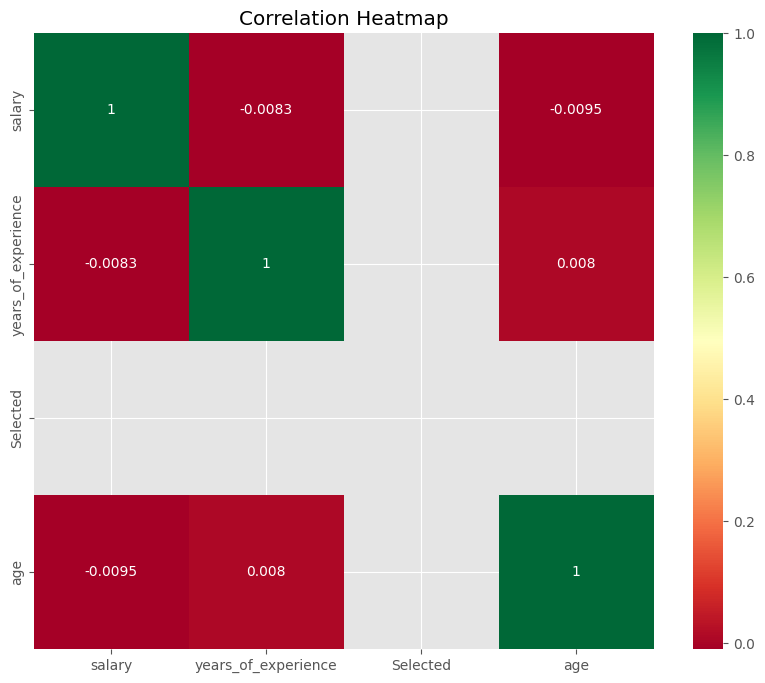
* Calculating average salary by city and visualizing using a bar plot
* Displaying the bar plot showing average salary by city



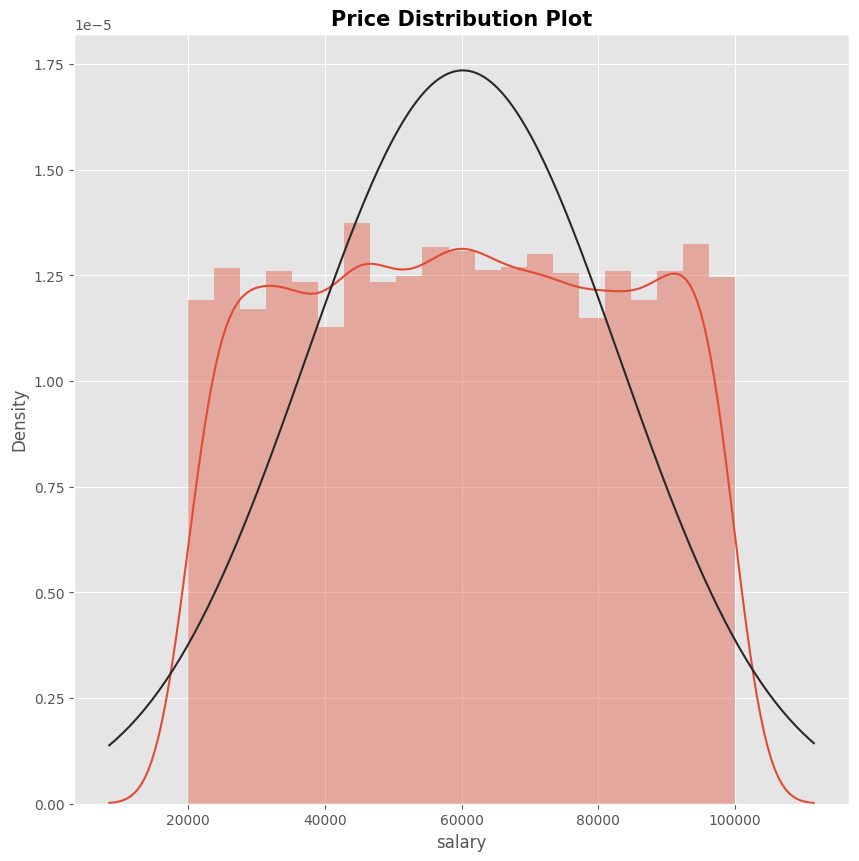
Conclusion: Displaying the top 10 cities with highest avg salary. The highest is hampur followed by kavali followed by Raichur

**Slide 12: Correlation Analysis**

* Calculating correlation matrix between numerical variables
* Visualizing correlation matrix using a heatmap
* Displaying the heatmap using plt.show()



**Conclusion**: Not much Correlation can be seen between the attributes

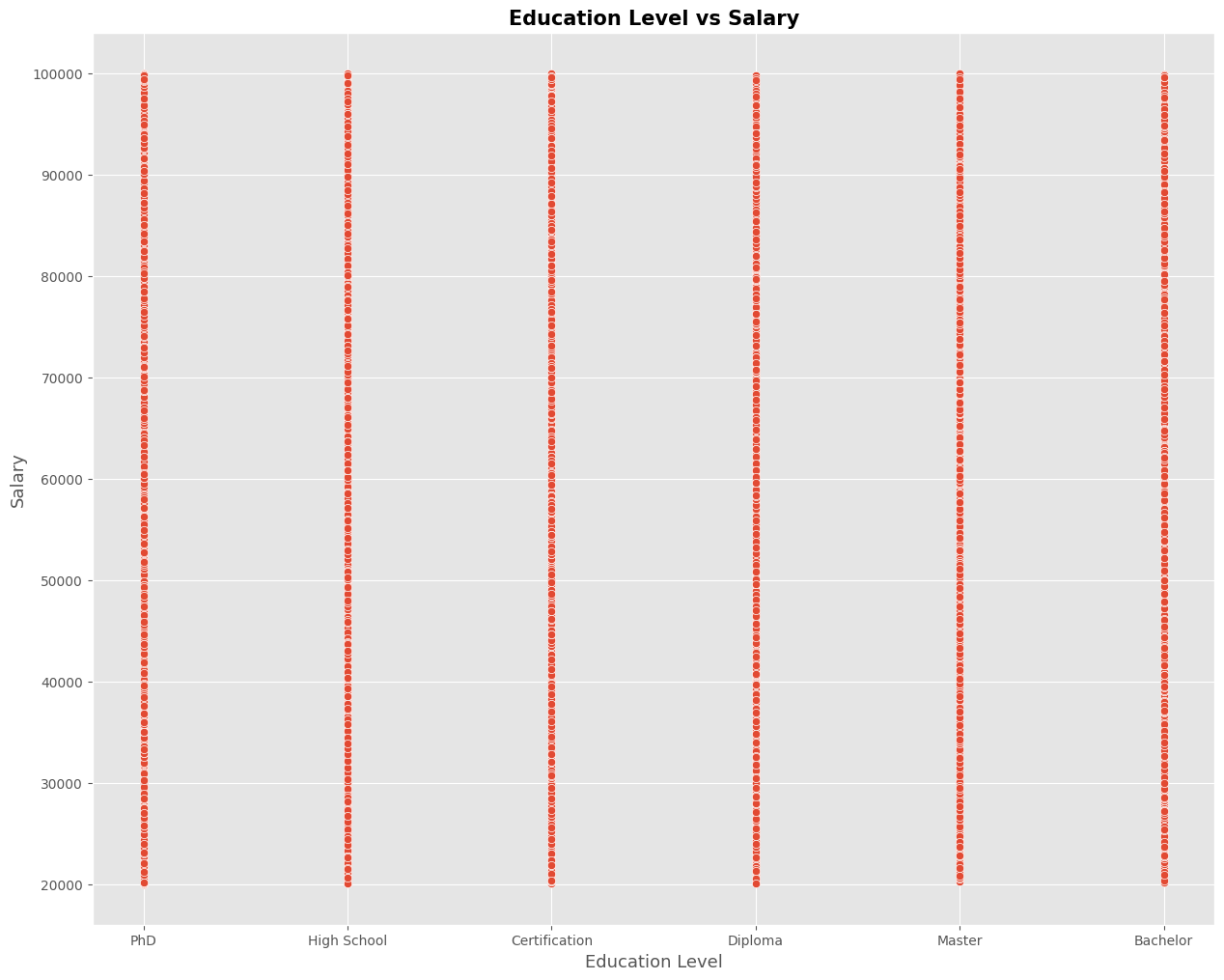


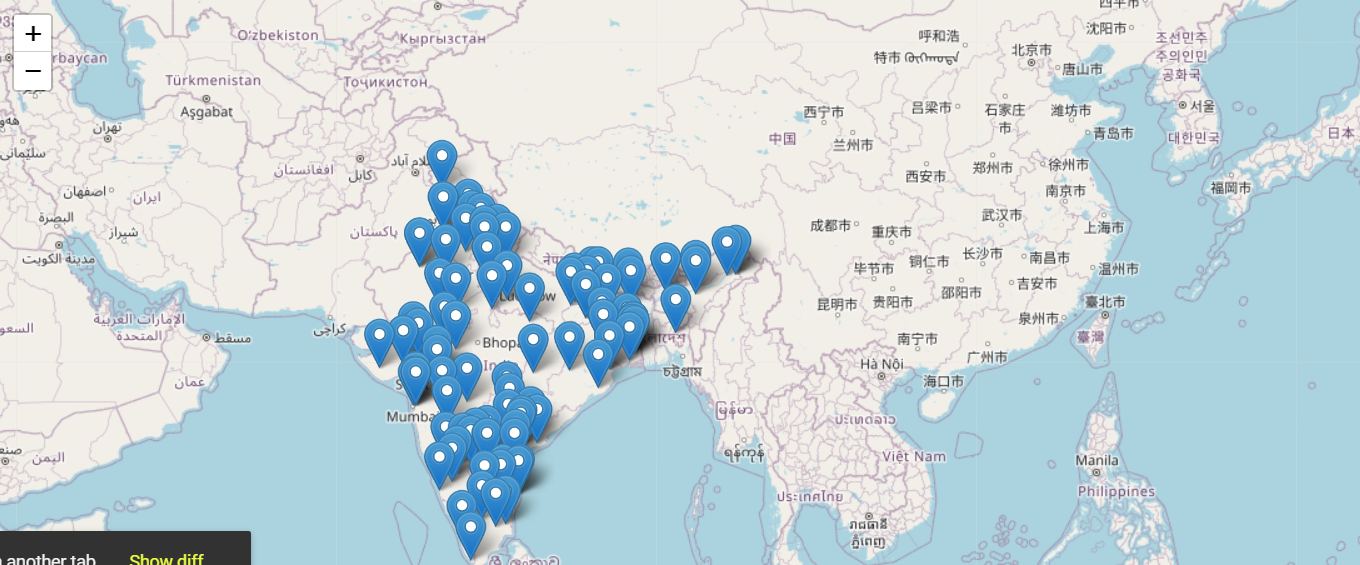
**Conclusion:** We can see that avg salary is around 60000 and there is not observed skwenss.

**Slide 13: Geocoding and Mapping**

* Extracting latitude and longitude coordinates for cities using Geopy
* Plotting cities on a map using Folium
* Displaying the map with markers for each city

**Slide 14: Plotting cities on map of India**

* Given the names of cities, we can map their latitude and longitude coordinates and display them on a map of India. Due to time constraints, only 100 cities were mapped.
* 

****