Week #2: Data Properties with Plots

Amir Fawwaz

June 8, 2023

1 Overview

In week #1, you have learnt on how to read and inspect *basic* properties of your data.

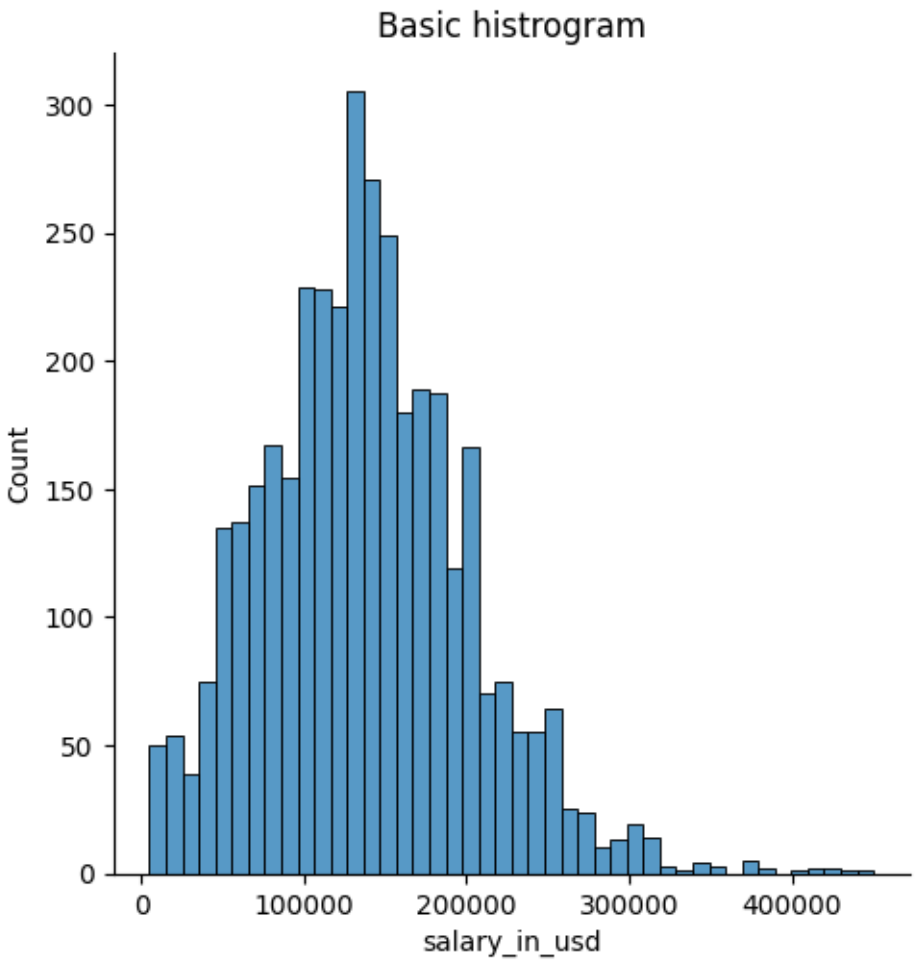
Now, let's explore about data distribution using *matplotlib* and *seaborn* package.

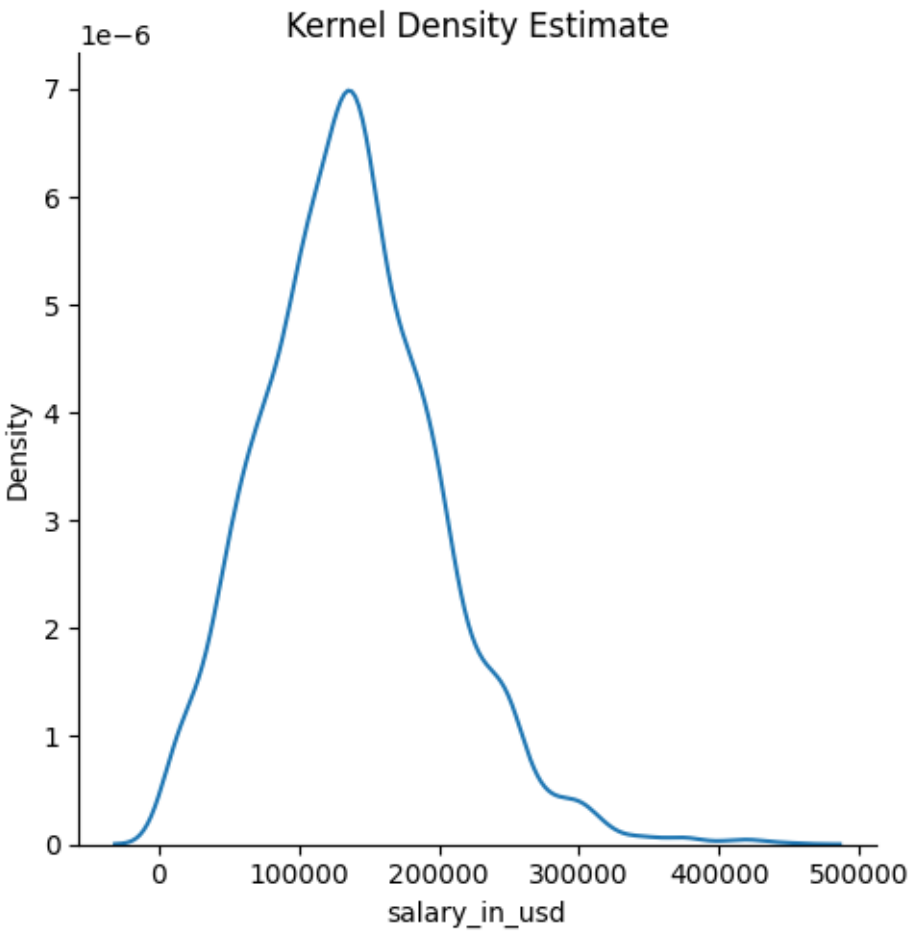
1.1 Sample

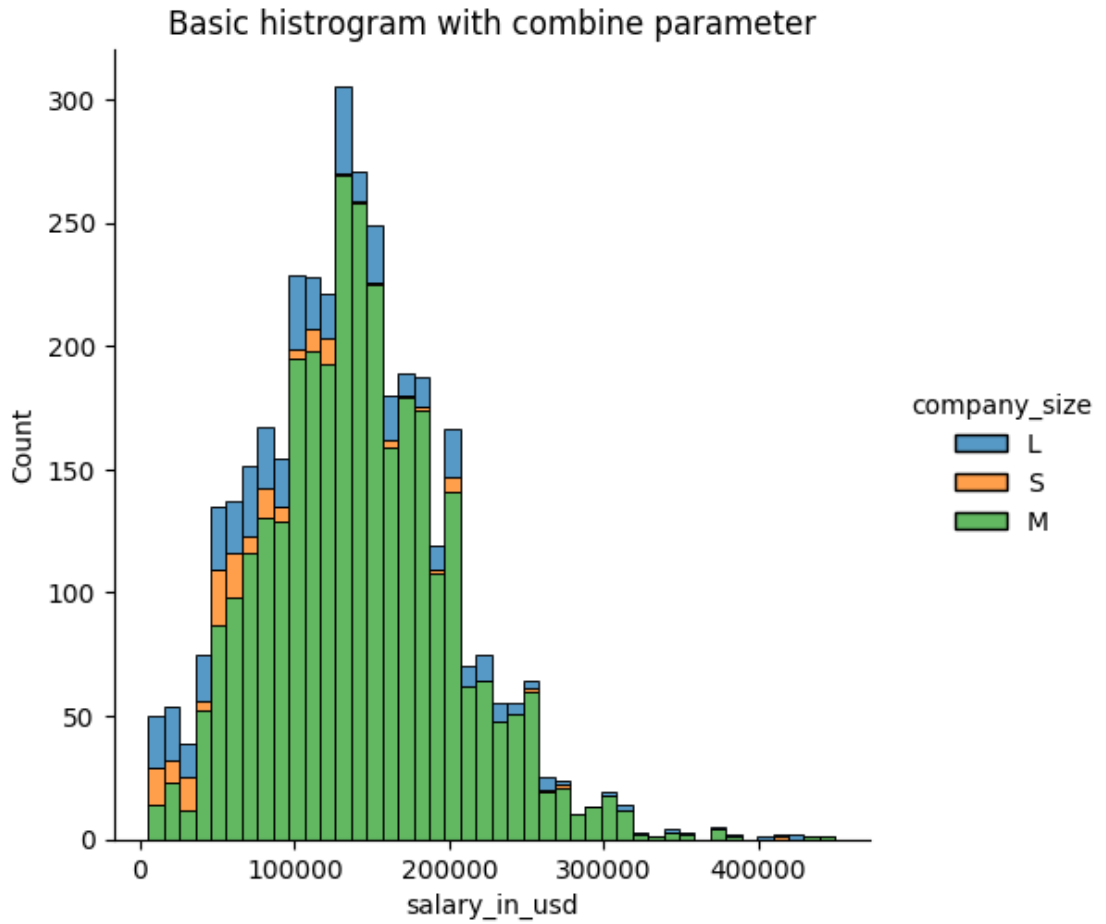
Image you have data tabulated like table below.

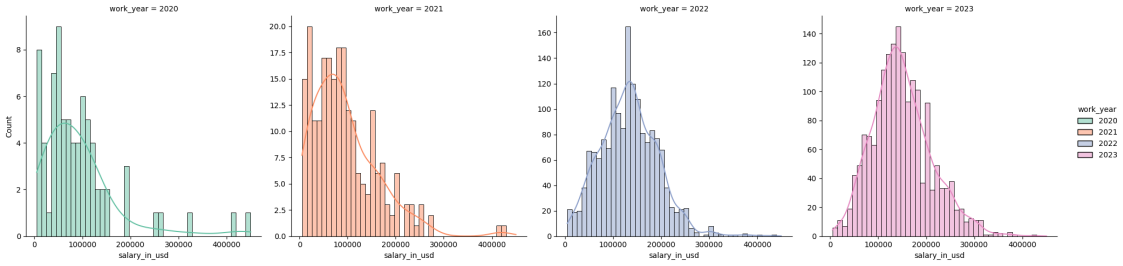
	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	employee_residence	remote_ratio	company_location	company_size
0	2023	SE	FT	Principal Data Scientist	80000	EUR	85847	ES	100	ES	L
1	2023	MI	CT	ML Engineer	30000	USD	30000	US	100	US	S
2	2023	MI	CT	ML Engineer	25500	USD	25500	US	100	US	S
3	2023	SE	FT	Data Scientist	175000	USD	175000	CA	100	CA	M
4	2023	SE	FT	Data Scientist	120000	USD	120000	CA	100	CA	M

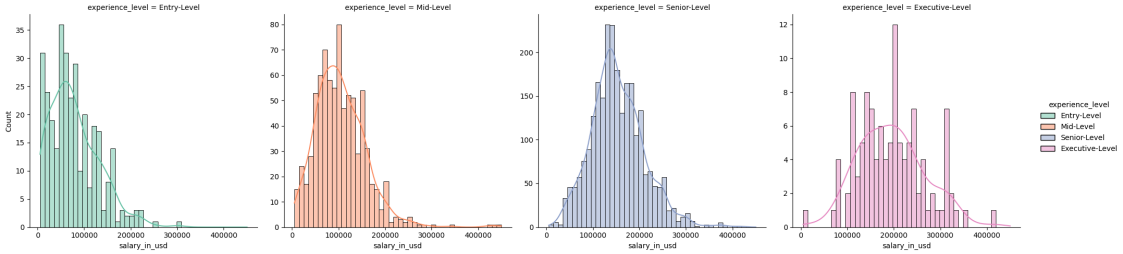
Now let's investigate salary distribution with histogram and density plot

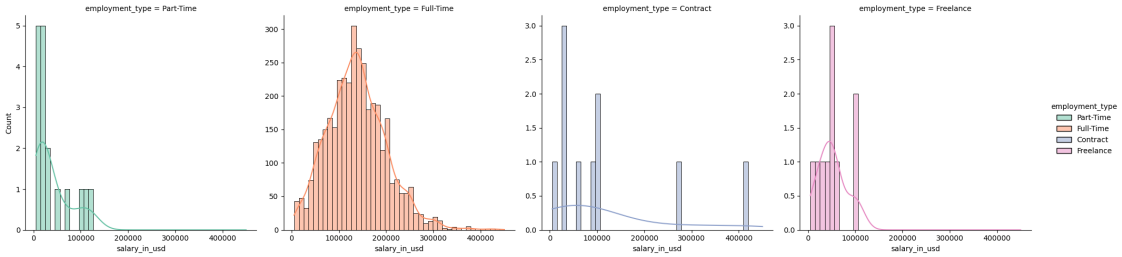

Basic histrogram
Count
300
250
200
150
100
50
0
0
100000
200000
300000
400000
salary_in_usd

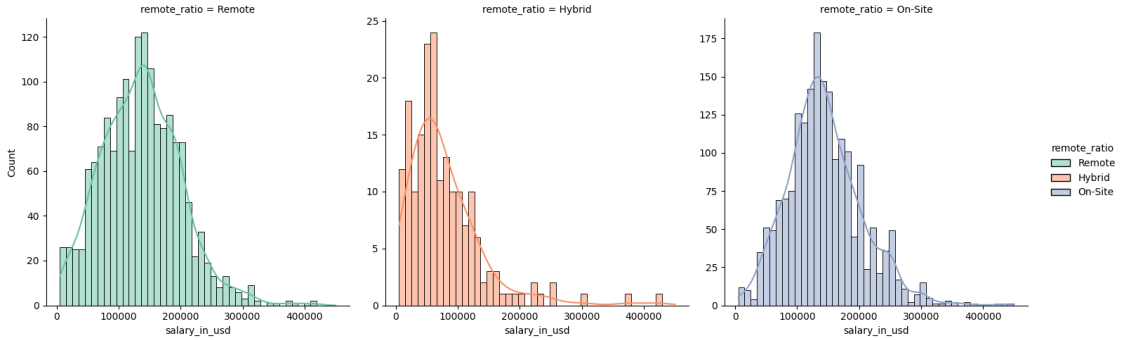

Kernel Density Estimate
1e−6
7
6
5
4
3
2
1
0
Density
0
100000
200000
300000
400000
500000
salary_in_usd

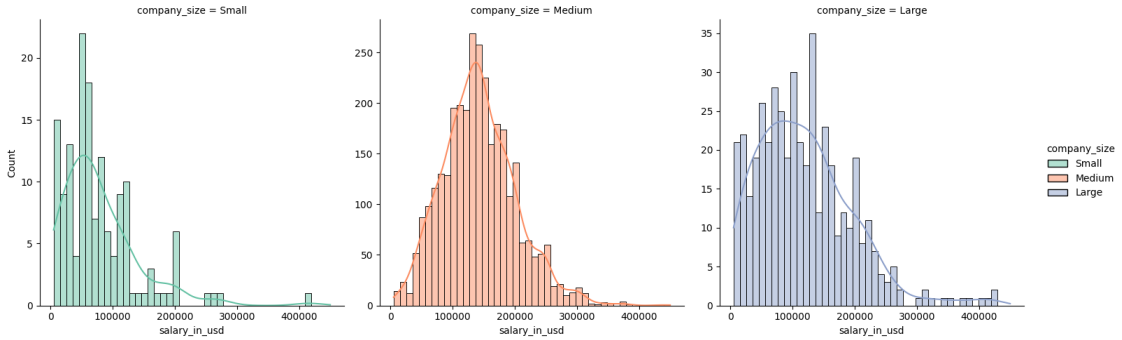

Basic histrogram with combine parameter
Count
salary_in_usd
company_size
L
S
M


work_year = 2020
work_year = 2021
work_year = 2022
work_year = 2023
Count
salary_in_usd
work_year
2020
2021
2022
2023


experience_level = Entry-Level
experience_level = Mid-Level
experience_level = Senior-Level
experience_level = Executive-Level
Count
salary_in_usd
experience_level
Entry-Level
Mid-Level
Senior-Level
Executive-Level

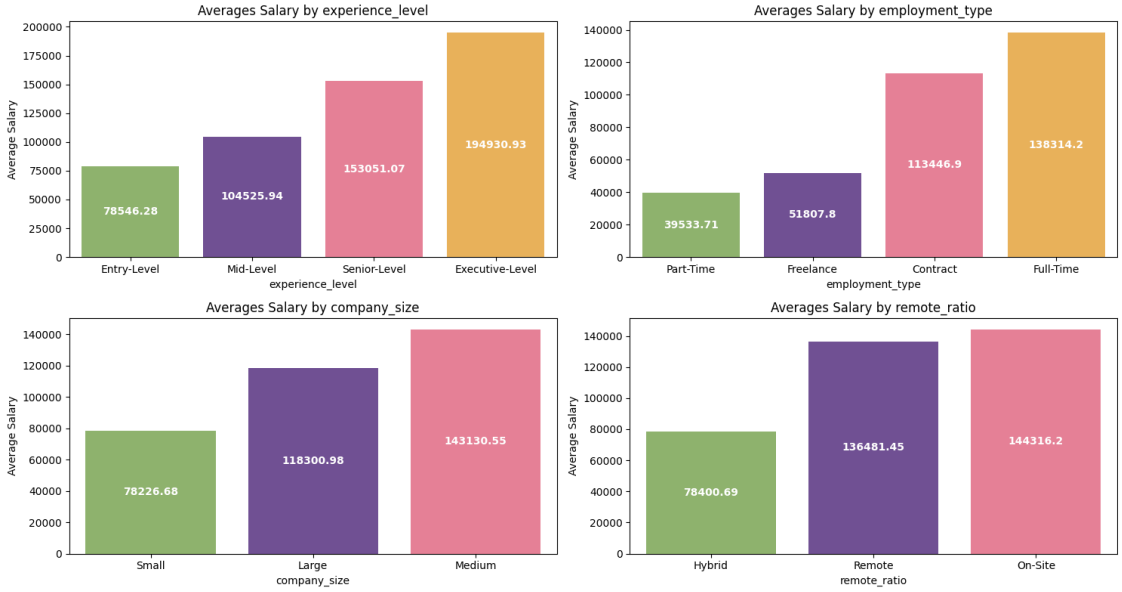




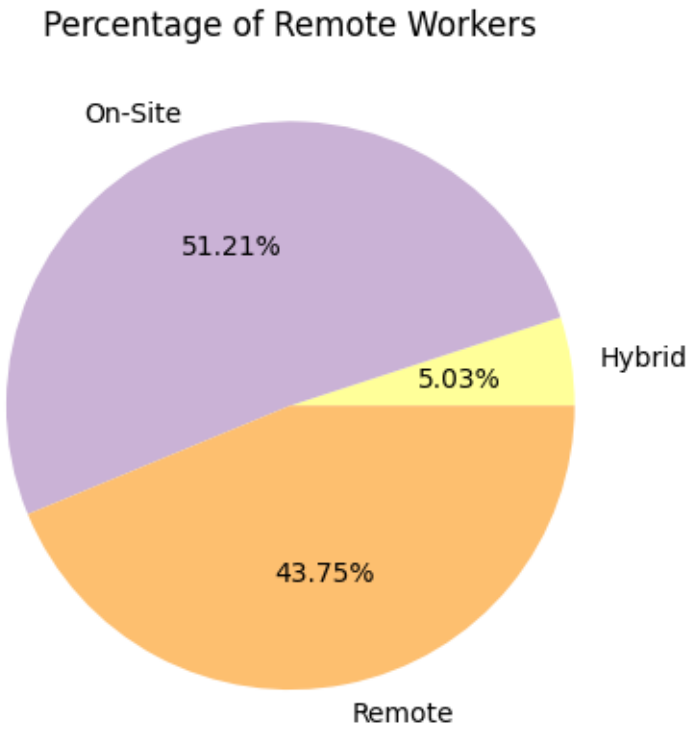


Note: What is density plot? What does it show us?

How about average salaries according to `company size` , `experience level` and others.



How many people working remotely compare on on-site?



1.2 Task 1

Library package to use:

- matplotlib
- seaborn
- plotly (optional, but you can try to explore)

Plot to explore:

- pie chart
- histogram
- density plot
- multiple plot

Data to use:

- Coffee Quality Data
- Bank Customer Churn
- Mushrooms images classification 215
- 5 Flower Types
- Headgear 20 classes-Image
- Chest CT-Scan
- Store Sales
- IceCube Experiment
- Cervical Spine Fracture
- Great Barrier Reef

  Note: Some of the dataset are very big (> 1 GB). Please download accordingly

1.3 Task 2

From Task of week #1 using the this data, do the following:

- seperate the image according to type
- plot suitable plot to describe the dataset (e.g. is the total number of image the same for both type? are all the image same size?
- add external images to the dataset & re-plot. Observed the data distribution.