**EDA (Importance of Histograms)**

Sometimes, do you often get confuse that when and how to use histograms while analysing the data.

Do you often use trial and error methods to see whether the presentation is up to your expectations via using univariate analysis (some other graphs) or via histograms.

This is because, you don’t know the exact meaning and use of histograms about how and when to use it.

Before, directly jumping the use case and meaning of histogram, we should know what is **univariate analysis, because from here you can get the clear picture, why, when and how to use histograms for analysis.**

|  |
| --- |
| Let us understand these concepts vis practical application in python with IRIS dataset.  Wondering – 1. Why IRIS dataset only.  2. Why univariate approach for understanding Histograms. |

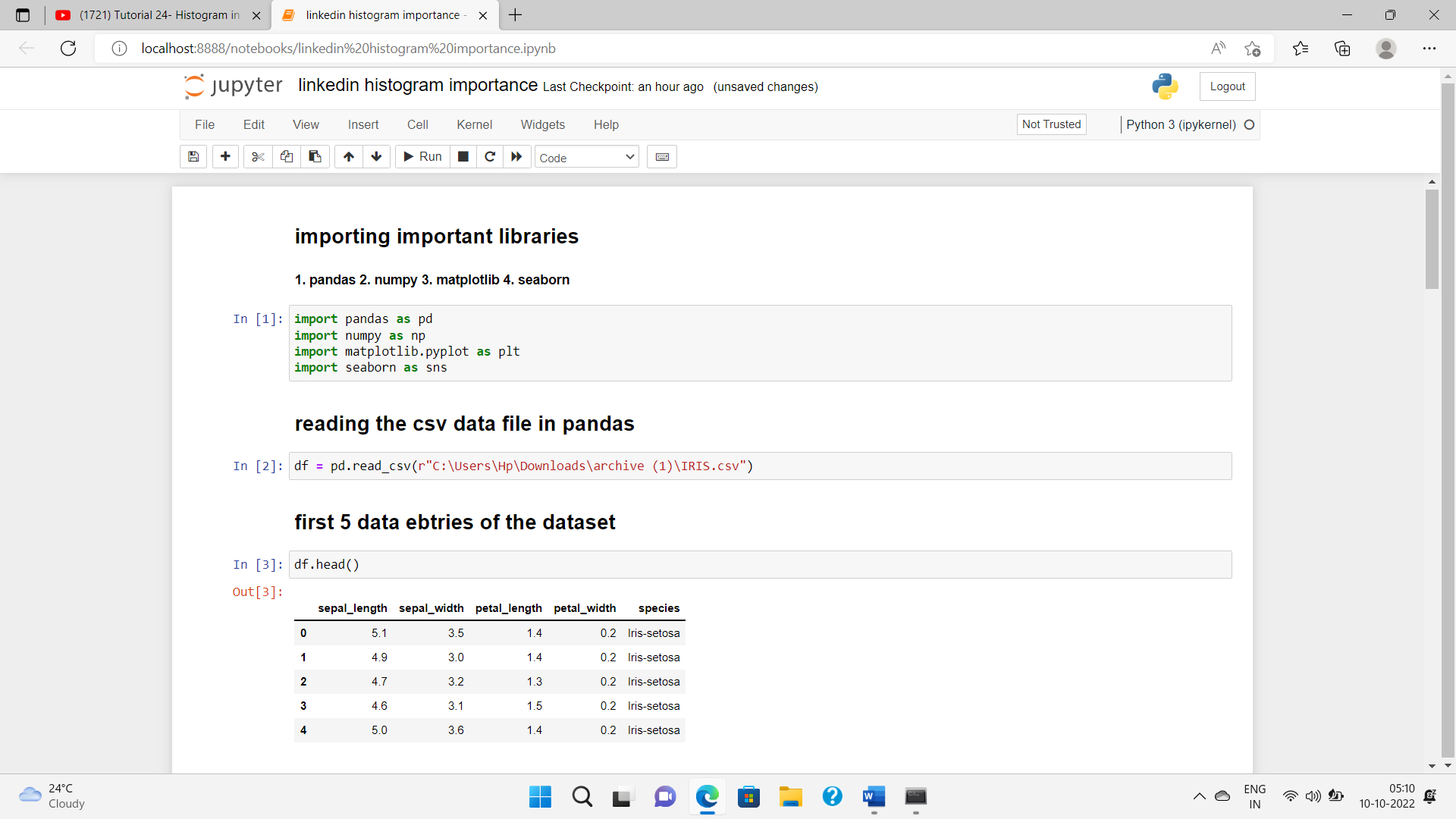
**Answer for 1 curios question.**

1. Small dataset
2. Have one categorical feature as SPECIES, perfect way to understand both of these concepts.

**Answer for 2 curios question.**

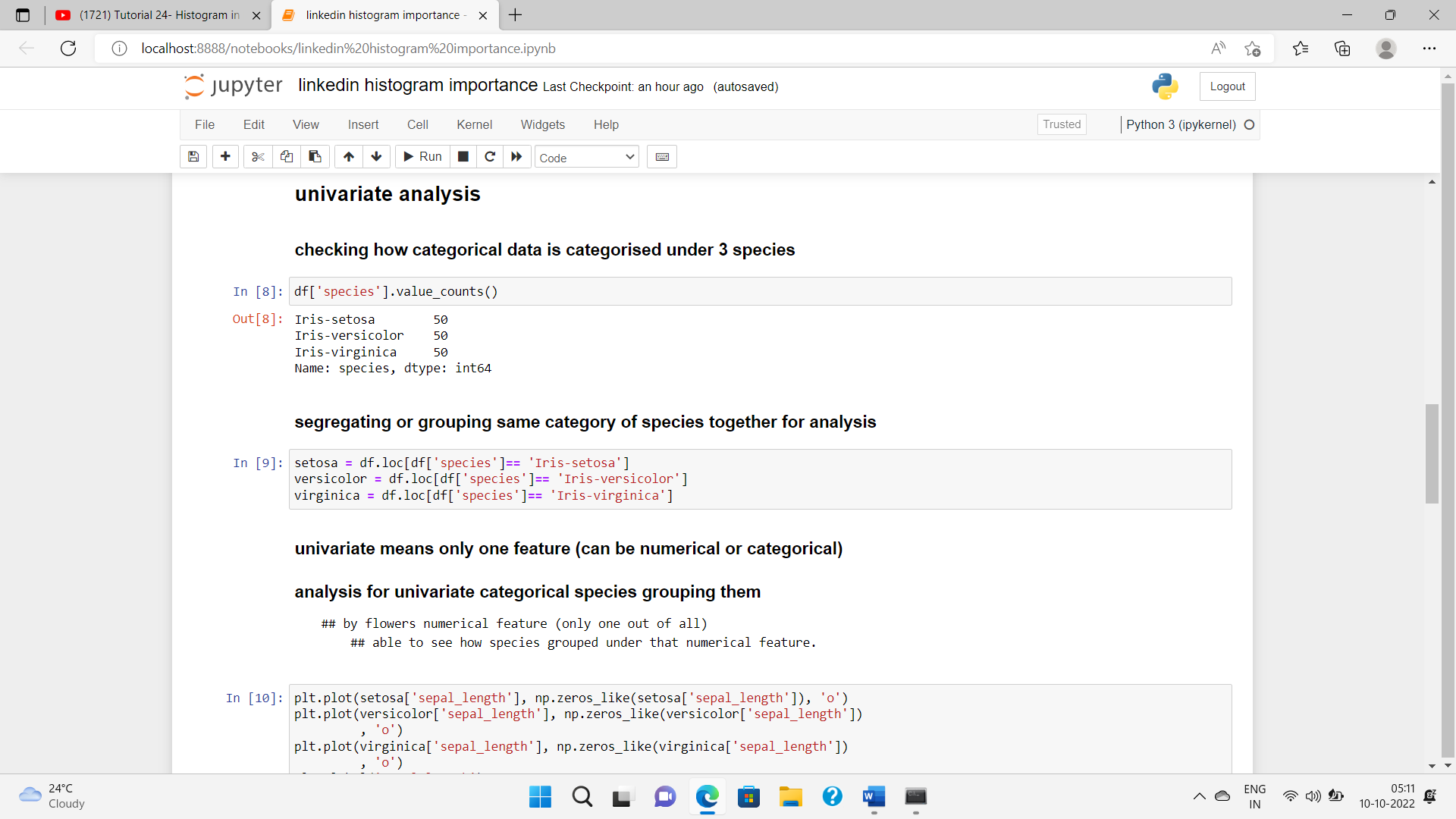
This is the one way for understanding the importance of histogram via univariate analysis.

I am sharing this. ipynb file (my working screenshots) with all of you, here I have applied the concepts practically in python.

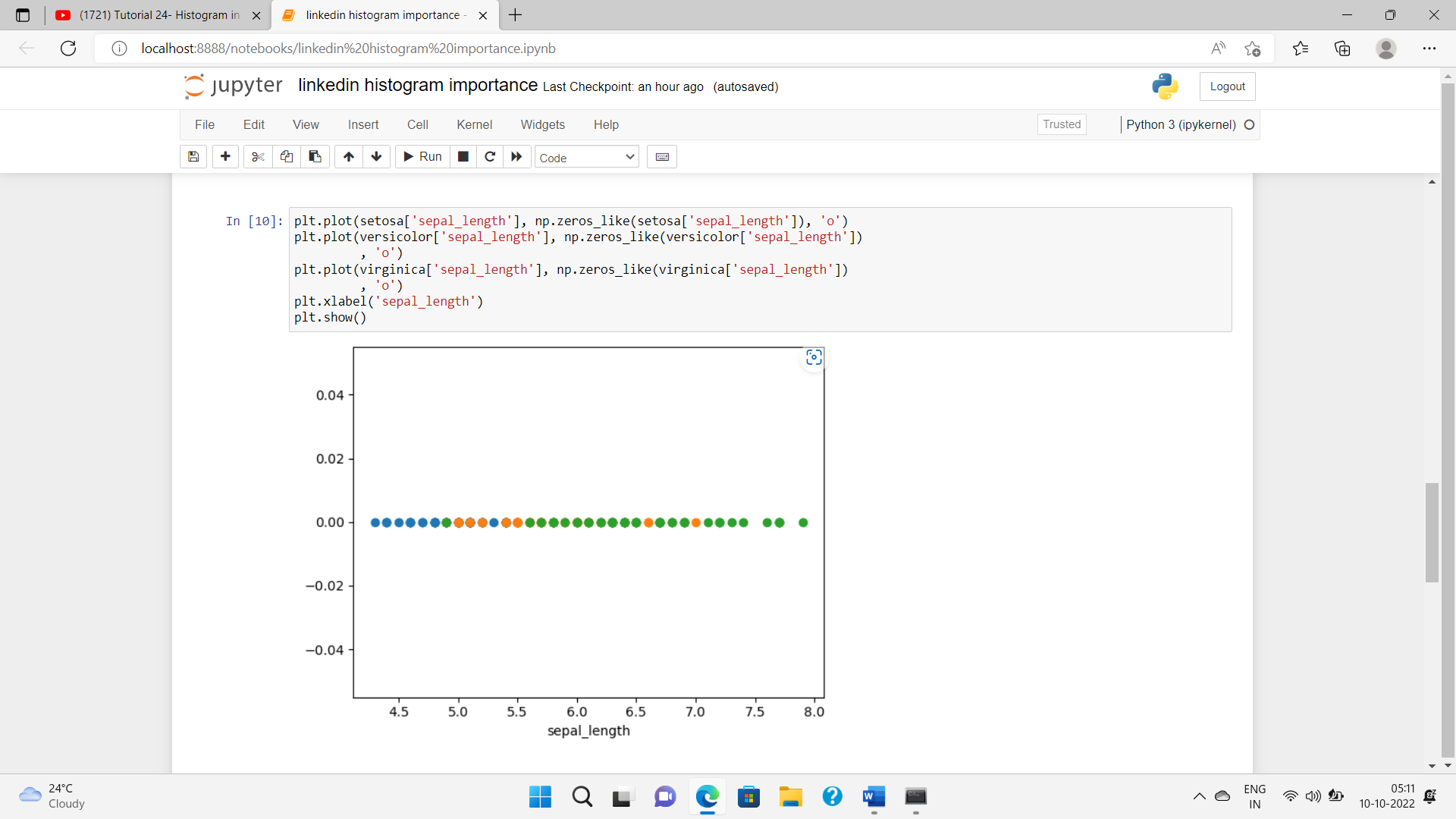


**Here, what we are trying to see that why histogram is so important for analysis.**

|  |
| --- |
| **In univariate analysis we first require analysis of only one feature (either numerical or categorical) and from our dataset what we require is how species (categorical feature) is categorised further into 3 different species.**  **For this, we have inbuilt function loc (location) in pandas for selecting the range of setosa, versicolor and virginica datapoints.** |



**Now, we need to plot the graph via using matplotlib in python.**



Here, analysis for sepal length (selected for univariate analysis), you can use any other single feature too on x – axis for analysing its behaviour for all species categories.

Remember – only one feature for analysis

|  |
| --- |
| **Yes, I got your question, what is represented on y-axis?**  **Answer – nothing, in fact a vertical line is drawn for y-axis to have all values zero.** |

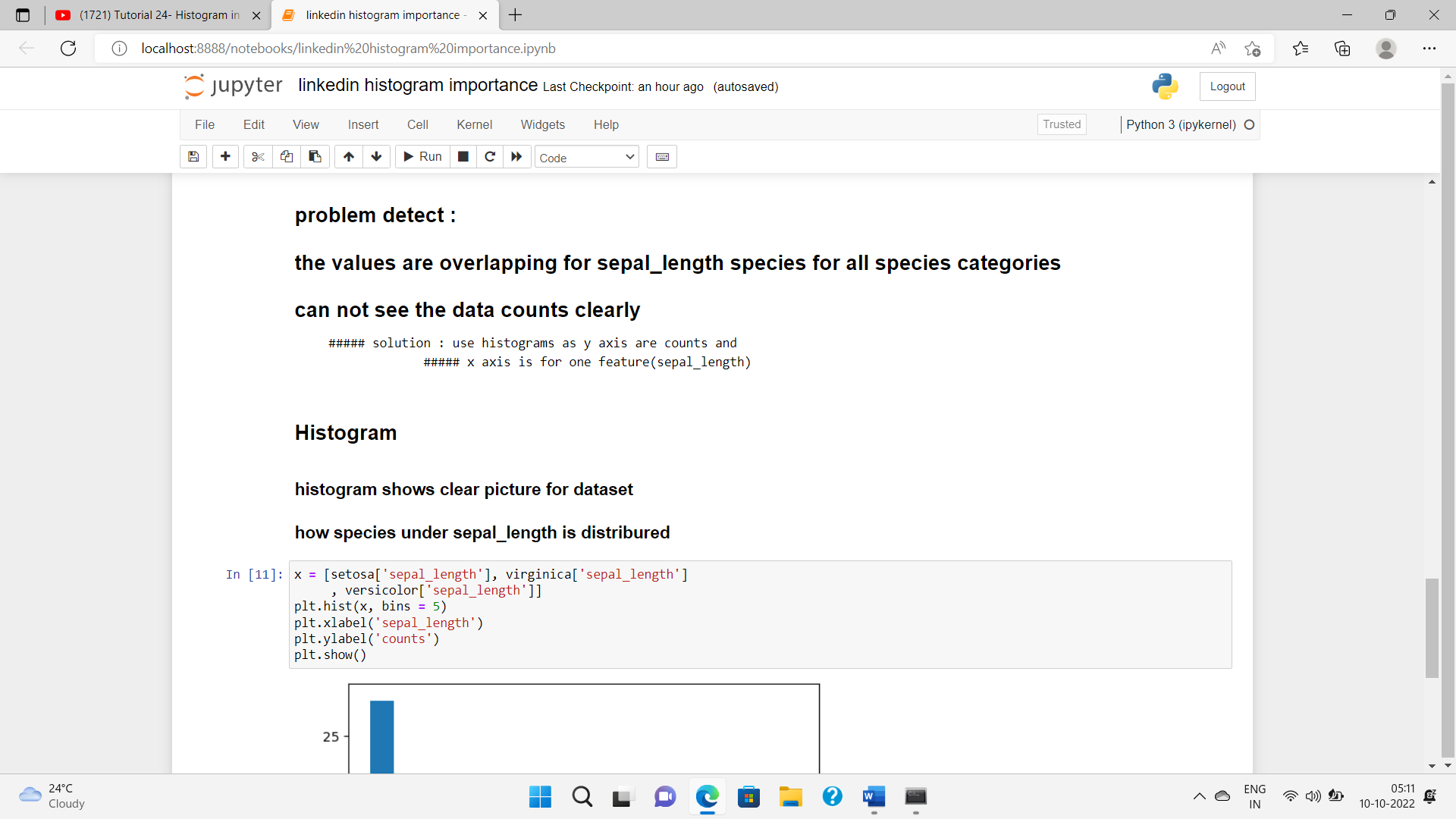
|  |
| --- |
| Analyse the graph  **Blue points are for setosa species for sepal length**  **Orange points are for versicolor for sepal length**  **Green points are for virginica for sepal length** |

Useful insights from graph:

1. The green points have outliers as range for more green points are within 5.5 – 8 but it has one datapoint in 5.0 range of sepal length.
2. Same for orange and blue datapoints, cannot obtain a clear presentation of data group.

***What do you think- are you able to make any useful insight from above univariate figure?***

***The classification of each species are overlapping, and difficult to analyze the datapoints segregation clearity. (no clear range is specified)***

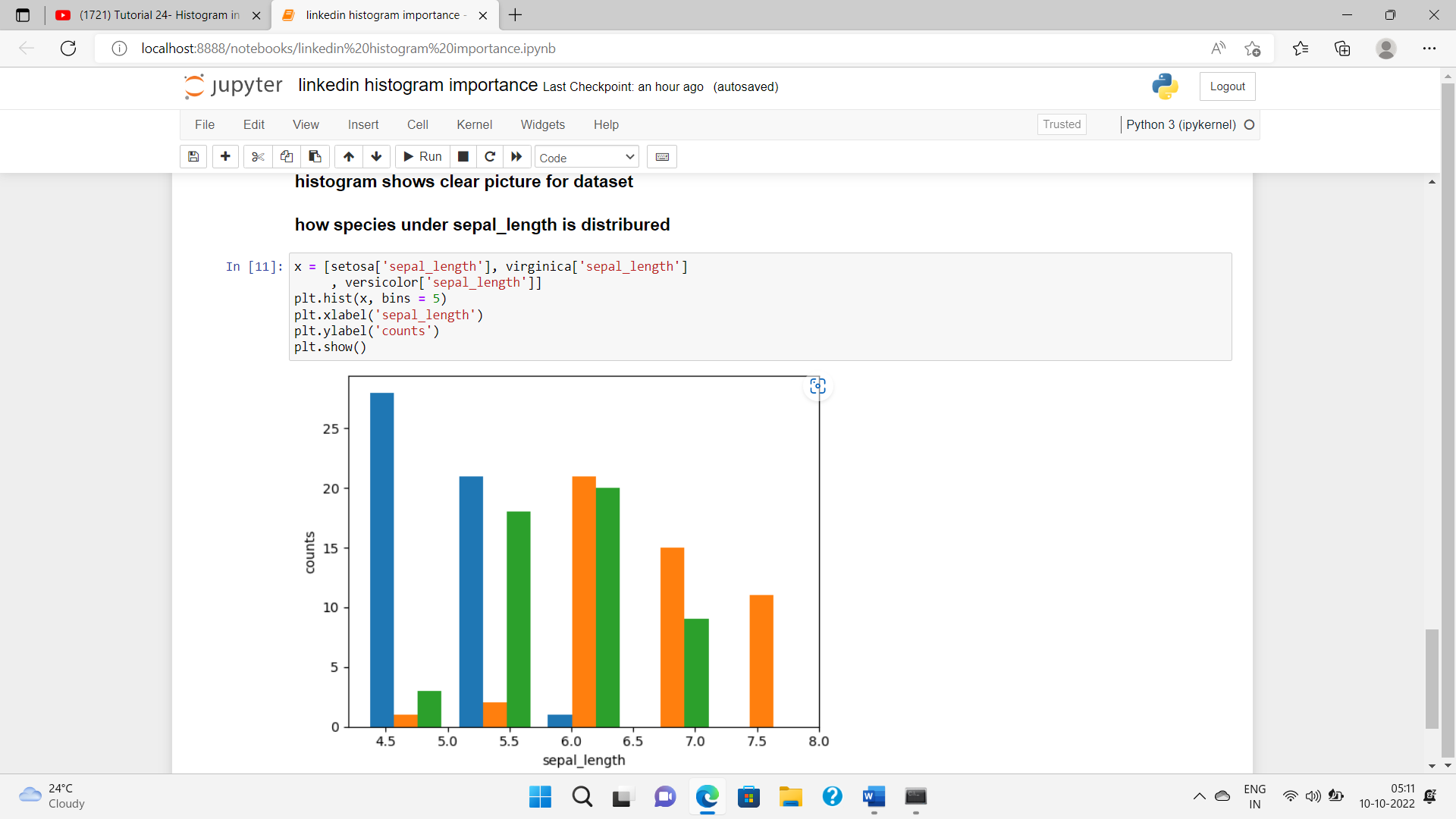


|  |
| --- |
| **Why need histogram**   1. **There are so many points falling in the (4.5 – 5.5 , say) range that we cannot figure out classification easily (univariate 1D analysis)** 2. **Impossible to count for each range clearly** 3. **Normally disrtributed data.** |

***How histogram helps***

1. ***Histogram y-axis is for counts and helps to count how many datapoints of which category of species is of sepal length (count of each range for that feature).***
2. ***For each range, the bar( rectangular box) will be created and from we can even analyse the proper classification , exact counts and where which species is highest and lowest.***
3. ***We can draw KDE or convert it into PDF and can use various libraries to plot histogram***

***Here, I have used matplotlib.***



Normally what we think earlier histogram is only helpful mainly for Checking how data is normally distributed or not.

**SUMMARY**

**But, there are other important points to consider like analysis considered over here.**

**Like:**

1. **Count**
2. **Comparison of datapoints**
3. **Clear representation of grouped data.**