AAI 614 - Data Science and its Applications
Graded Assessment: Project 1

Linear Regression Analysis

Mostafa Zeinalabidine – 201201495

The objective of this project was to explore the 1994 US Census dataset and perform an exploratory data analysis as preparation for linear regression training.

In addition to data cleaning and preparation, the tasks focused on uncovering patterns in features related to income, demographics, education, and workhours. The tasks also focused on understanding variable relationships and correlations.

Notebook link:

https://github.com/mostafa-zea/AAI614_-zeinalabidine-/blob/main/AAI_614O_Project1_Mostafa_Z.ipynb

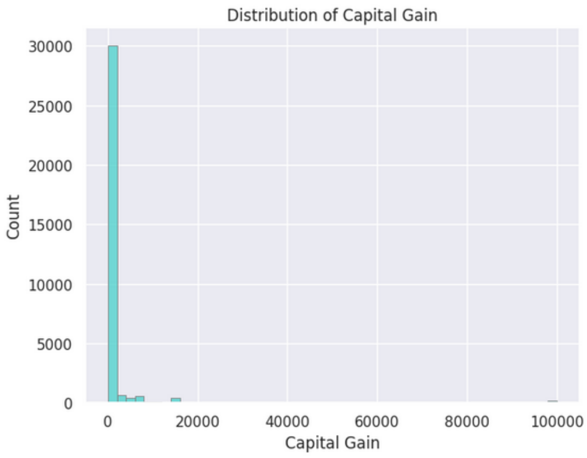
Dataset source and link:

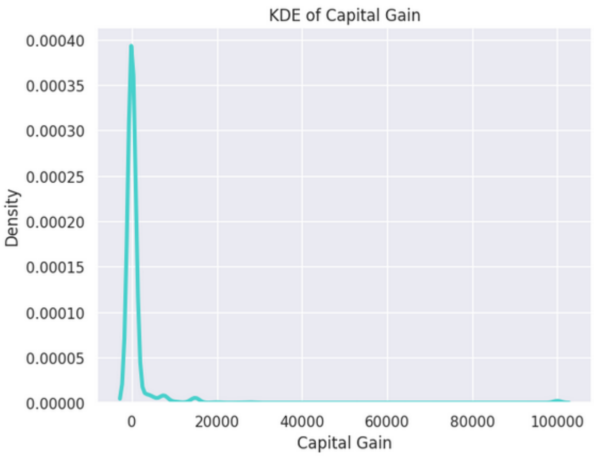
UCI Adult dataset

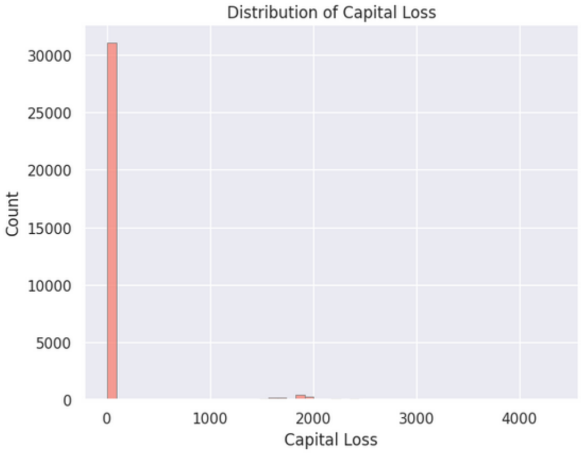
https://archive.ics.uci.edu/dataset/2/adult

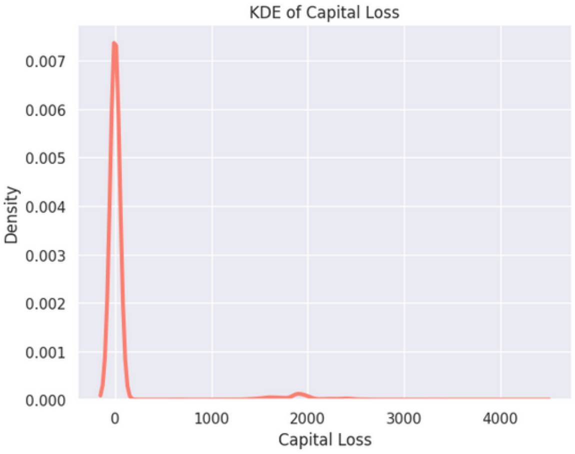
Data Preparation and Cleaning:

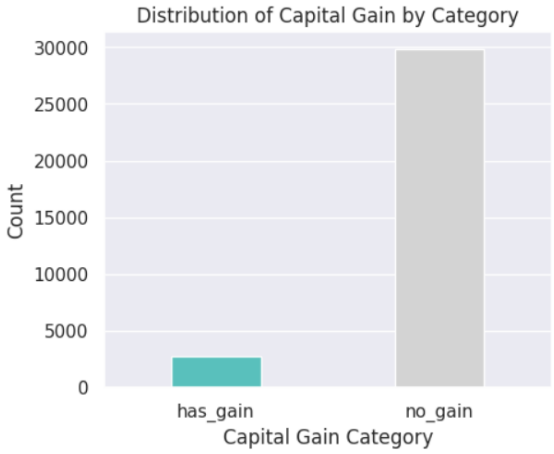
- Missing values which were represented as `“ ?”` in the data file were replaced with `np.nan`
- `skipinitialspace=True` was used to handle spaces after commas in the data file
- Missing values appeared in `workclass, occupation, and native_country`
- `capital_gain` and `capital_loss` had extremely right-skewed distributions, with most values being zero and few non-zero values. It was better to transform `capital_gain` and `capital_loss` into categorical binary features (`has_gain, no_gain, has_loss, no_loss`).

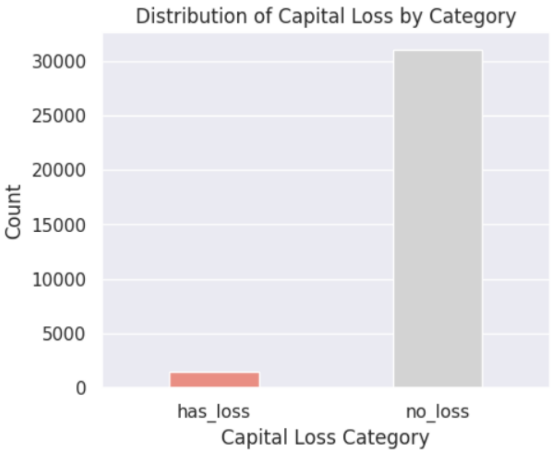








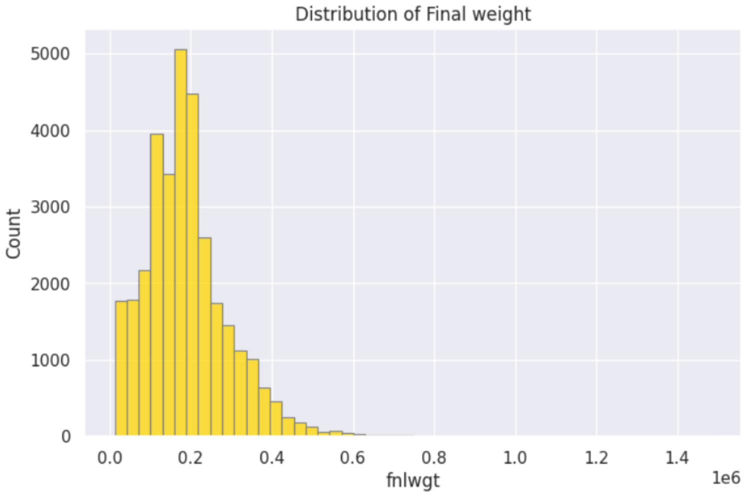


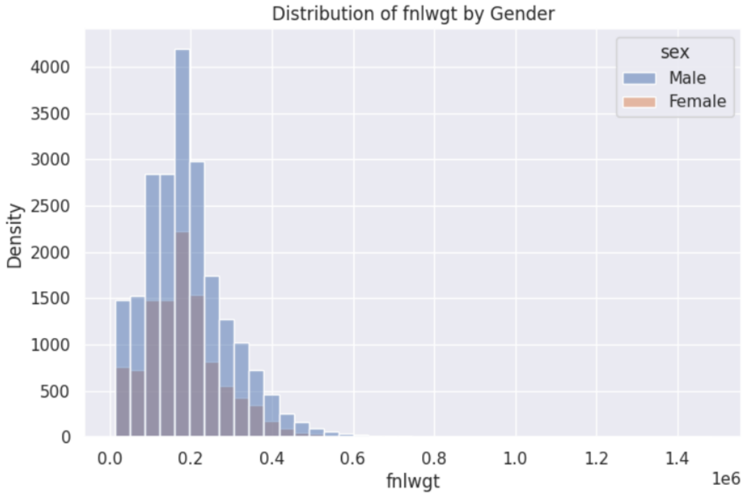


Exploratory Data Analysis and Correlation:

fnlwgt distribution:

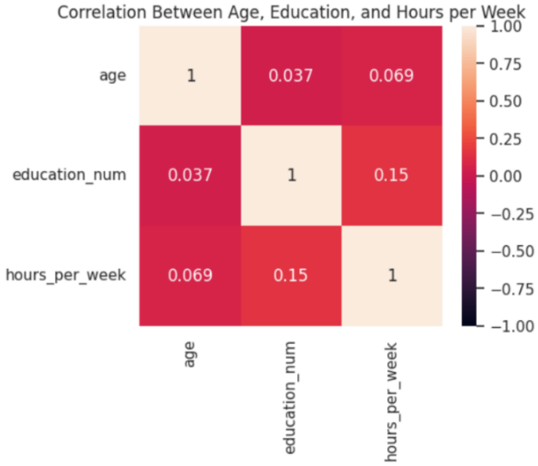
- Right-skewed distribution, most values concentrated between 150,000–250,000.
- Outliers were replaced with NaN (top 1% which were values closer to 1,000,000 far outside the typical range).
- Gender comparison showed almost identical distribution shapes but higher count for males. Therefore, the key difference in `fnlwgt` distribution between males and females was frequency and not trend.





Correlation

- Correlation analysis focused on `age, education_num, hours_per_week`:
- `education_num` and `hours_per_week` have the strongest positive correlation, (0.15). This suggests that individuals with more education tend to work slightly more hours per week.
- `age` and `hours_per_week` have a weak positive correlation (0.069). This suggests a minimal relationship between age and work hours per week.
- `age` and `education_num` are almost uncorrelated (0.037).
- None of the variables are strongly correlated (all 3three are closer to 0 than +1).



Pearson Correlation/ Statistical Significance

- Pearson Correlation showed significant p-values where expected.
- Testing the correlation between `education_num` and `hours_per_week` using Pearsons correlation test showed statistical significance.
- Testing the correlation between `education_num` and `age` for males using Pearsons correlation test showed that there is a weak but statistically significant tendency for `education_num` to increase with `age`.
- Testing the correlation between `education_num` and `age` for females using Pearsons correlation test showed that there is a no correlation and no statistical significance between `education_num` and `age`. This was expected since the dataset is from 1994, and might reflect historical or societal trends of the time.

Conclusion

The correlation and EDA performed in this project helped clean and transform the data, reveal patterns and trends, and identify important relationships between features. Which is an important step in preparation for linear regression training in the next phase using variables like education, working hours per week, and demographic factors to estimate income category.