EDA

Beibei Du, Wafiakmal Miftah, Suzanna Thompson, Alisa Tian

2022-10-19

IDS702 Final Project EDA

Team member: Beibei(Bella) Du, Wafiakmal Miftah, Suzanna Thompson, Wenjing Tian

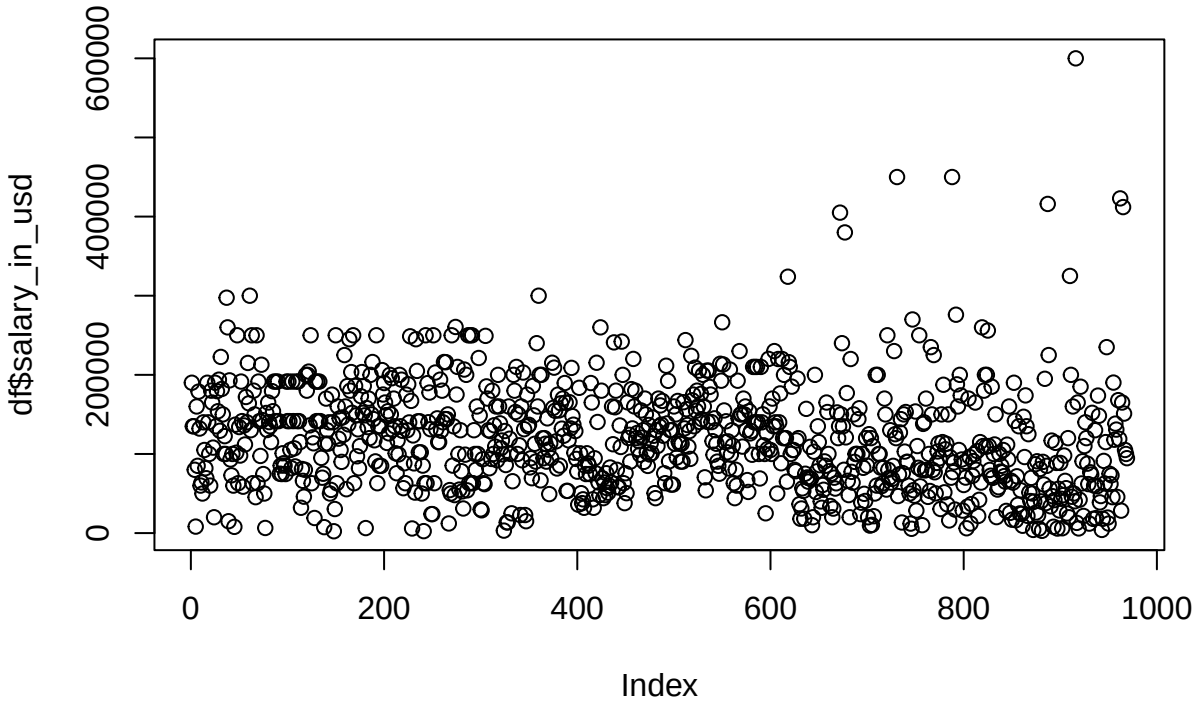
For part 1, we will conduct exploratory data analysis on our selected dataset.

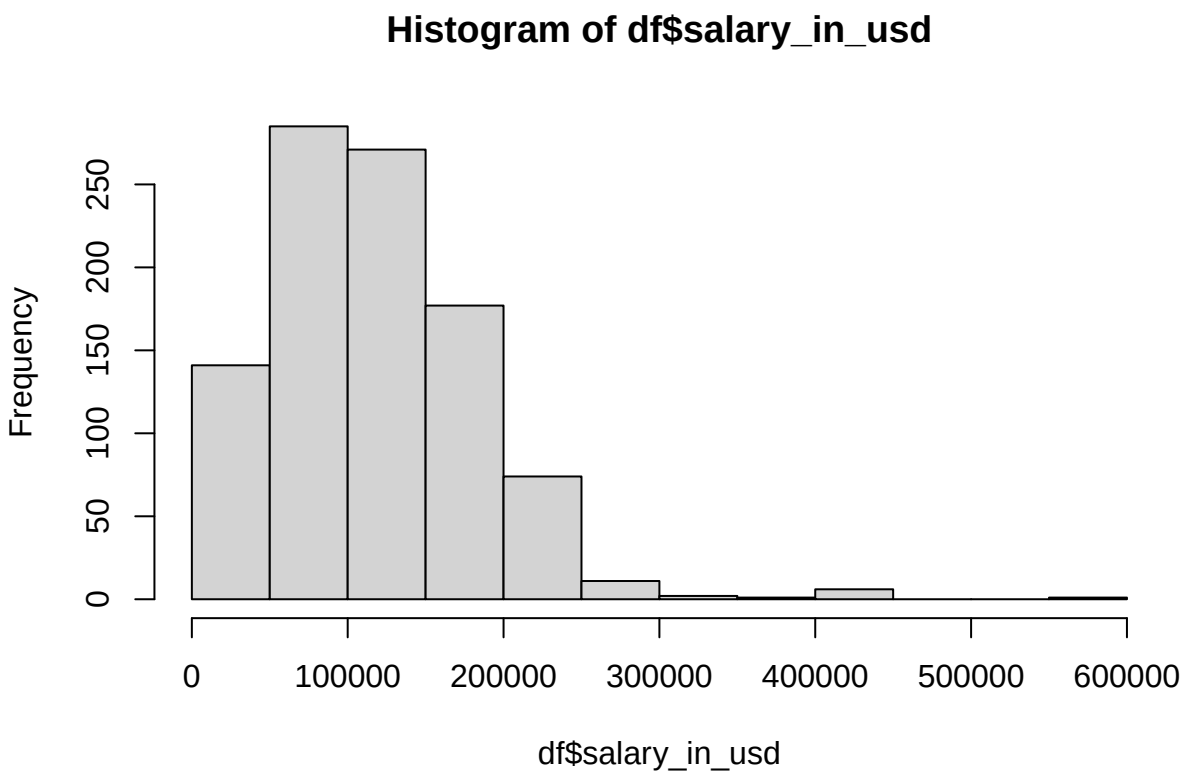
1. Data Overview

The data used in this project comes from Kaggle.

The link to the original post: Data Science Job Salaries

work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	employee_residence	remote_ratio	company_location	company_size
2022	SE	FT	Data Architect	190000	USD	190000	US	100	US	M
2022	SE	FT	Data Architect	135000	USD	135000	US	100	US	M
2022	SE	FT	Data Engineer	135000	USD	135000	US	100	US	M
2022	SE	FT	Data Engineer	80000	USD	80000	US	100	US	M
2022	EN	FT	BI Data Analyst	633000	INR	8191	IN	100	IN	M
2022	SE	FT	Data Engineer	160000	USD	160000	US	100	US	M


df$salary_in_usd
0
200000
400000
600000
0
200
400
600
800
1000
Index



1.2 Two proposed research questions

1. Which factors are associated with an increase in salary for data science jobs? (Continuous outcome)
2. How do company size, company location, employment type, employee residence and job title affected the remote work ratio of a data scientist? (Discrete outcome)

2. Primary relationship of interest

Present descriptive statistics and exploratory plots in whichever format you think is best (tables, figures) for your primary relationship of interest (dependent variable and primary independent variable, if applicable). Describe your findings.

2.1. Descriptive stats and plots Answering question 1

2.2. Descriptive stats and plots Answering question 2

3. Other characteristics

Briefly describe other variables in the data. If there are many, do not list them all. Rather, describe the types of variables that are present (e.g., "demographic information").

4. Potential challenges

Describe aspects of the data that may present challenges in the modeling stage. For example, might certain categorical variables need to be collapsed? Is there a lot of missingness? Could the size of the dataset present model selection challenges?