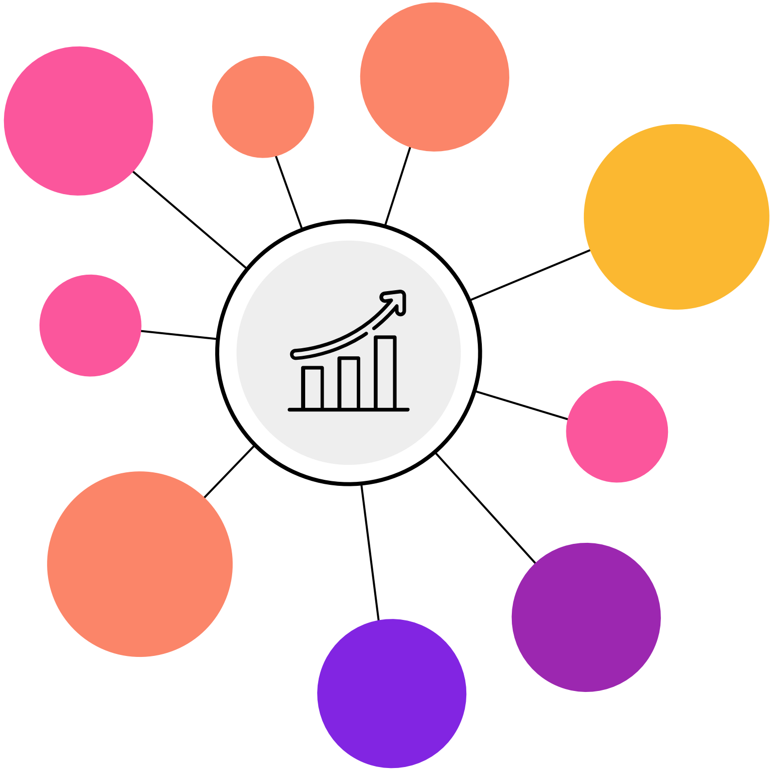


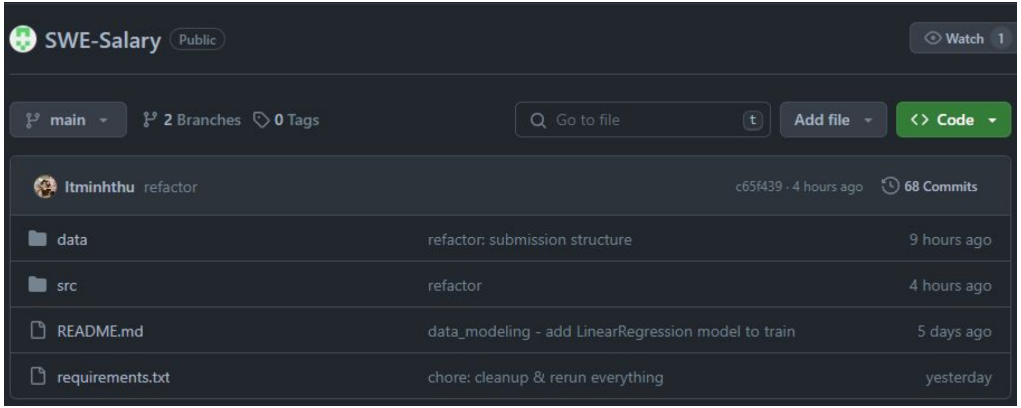
Software Engineer Salary

21KHDL1 - Group 7

Group Information

Members:
21127278 - Nguyen Trong Hieu
21127365 - Phan Phuoc Hai Nam
21127697 - Le Thi Minh Thu

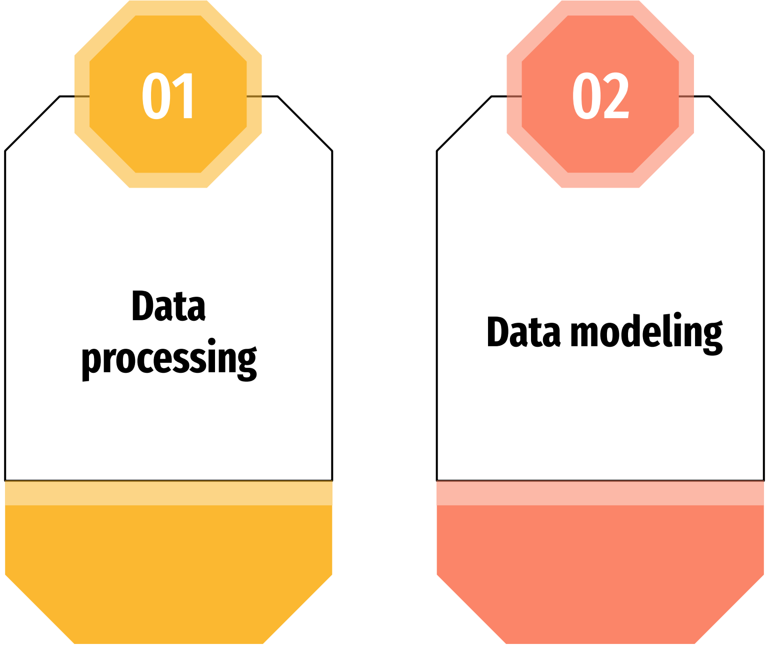
Github: SWE-Salary (68 commits)



ltminhthu refactor c65f439 · 4 hours ago 68 Commits

data	refactor: submission structure	9 hours ago
src	refactor	4 hours ago
README.md	data_modeling - add LinearRegression model to train	5 days ago
requirements.txt	chore: cleanup & rerun everything	yesterday

Software Engineer Salary



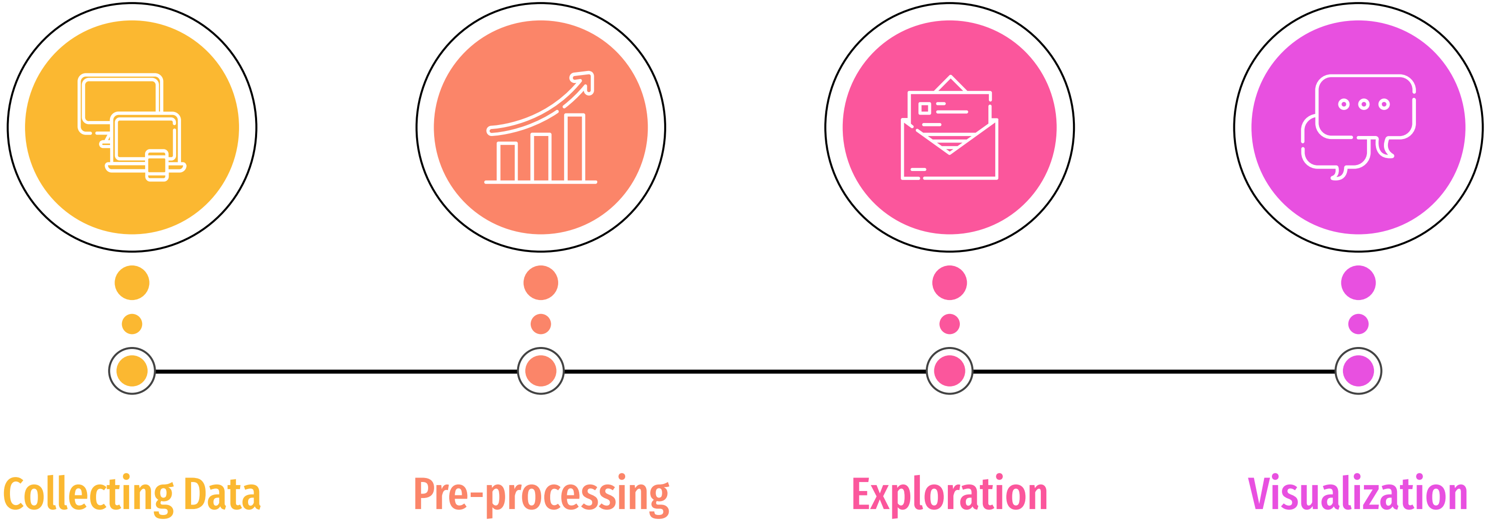
01

Data processing

02

Data modeling

Software Engineer Salary



Collecting Data

Pre-processing

Exploration

Visualization

Data Collection

01

Define the needed features

02

Find resources (APIs, web or static data)

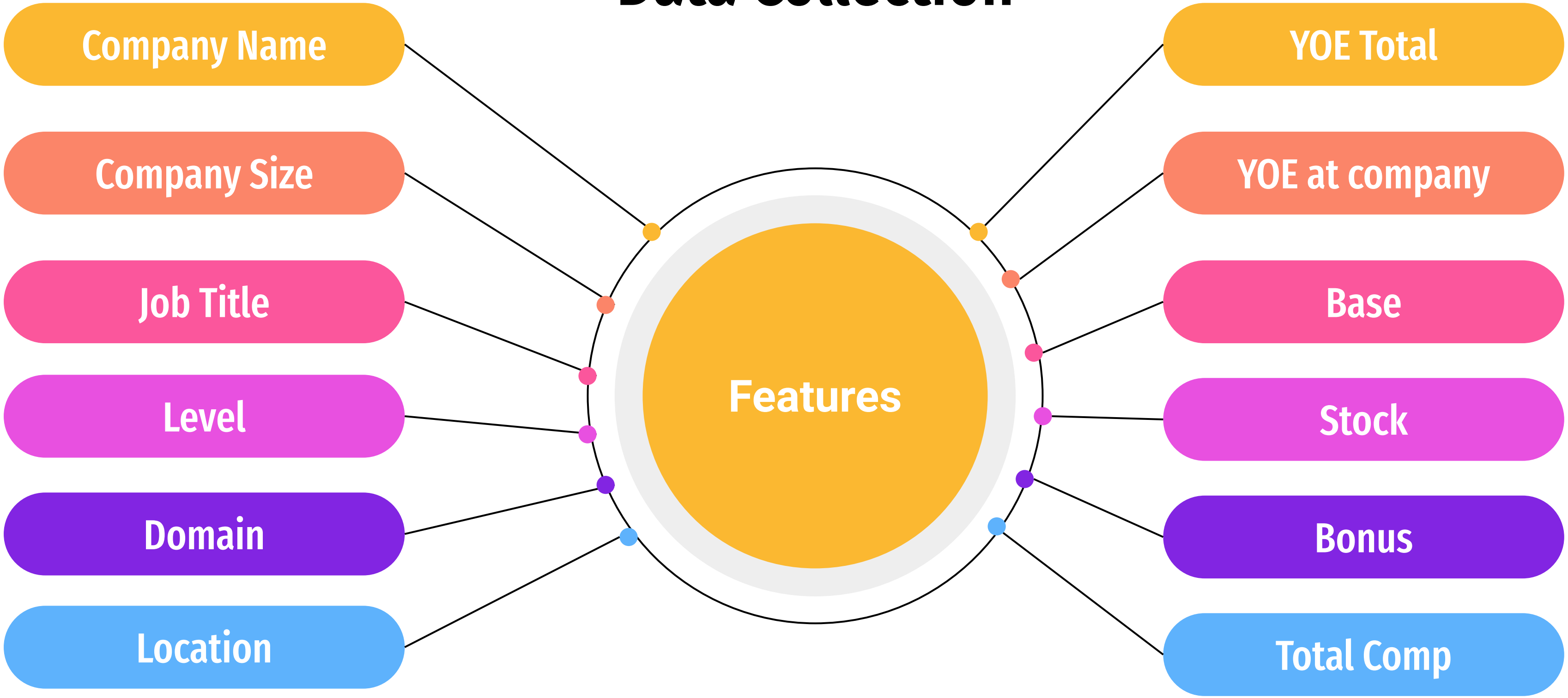
03

Locate the needed data

04

Get the data

Data Collection



Data Collection

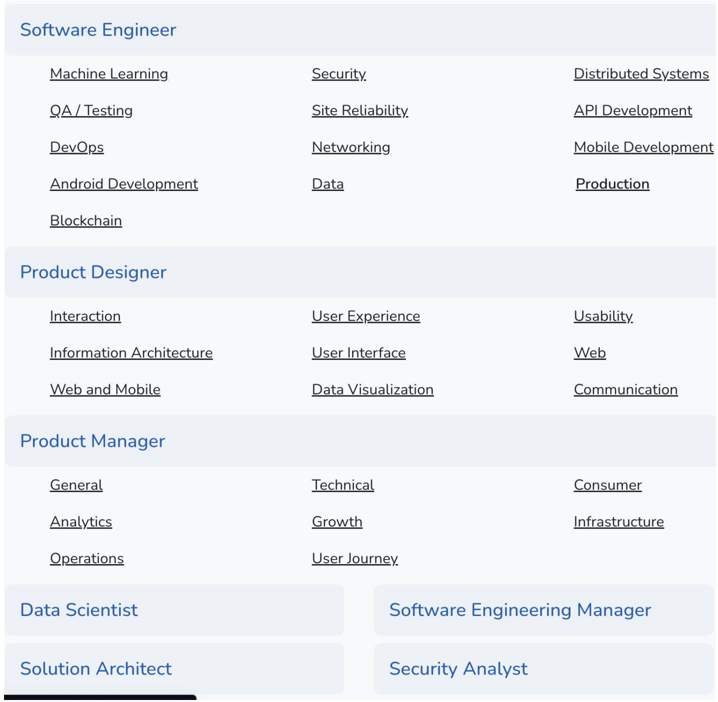
Resources

All the data is collected from levels.fyi:

- In 2022, levels.fyi had collected over 150.000 salary submissions.
- Their data comes from companies around the world.
- The information is well-defined and has a clear structure.

Data Collection

Resources



Software Engineer

- Machine Learning
- QA / Testing
- DevOps
- Android Development
- Blockchain
- Security
- Site Reliability
- Networking
- Data
- Distributed Systems
- API Development
- Mobile Development
- Production

Product Designer

- Interaction
- Information Architecture
- Web and Mobile
- User Experience
- User Interface
- Data Visualization
- Usability
- Web
- Communication

Product Manager

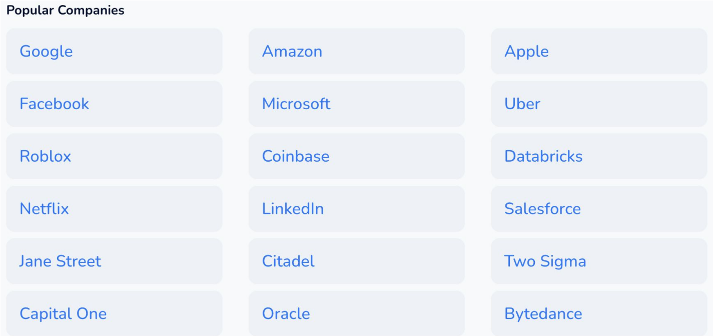
- General
- Analytics
- Operations
- Technical
- Growth
- User Journey
- Consumer
- Infrastructure

Data Scientist

Software Engineering Manager

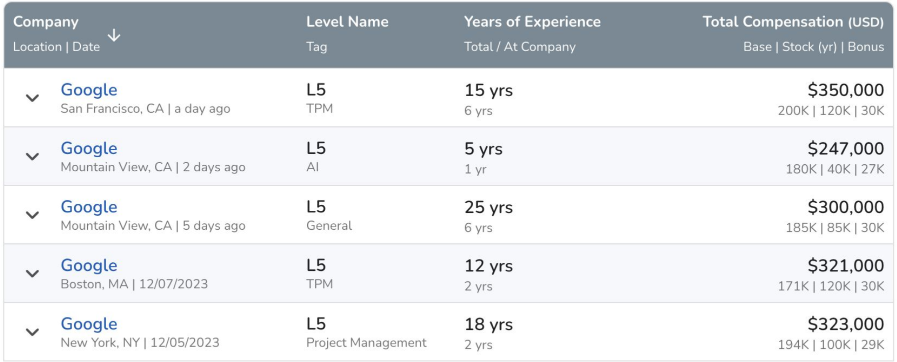
Solution Architect

Security Analyst



Popular Companies

Google	Amazon	Apple
Facebook	Microsoft	Uber
Roblox	Coinbase	Databricks
Netflix	LinkedIn	Salesforce
Jane Street	Citadel	Two Sigma
Capital One	Oracle	Bytedance



Company Location | Date	Level Name Tag	Years of Experience Total / At Company	Total Compensation (USD) Base | Stock (yr) | Bonus
Google San Francisco, CA | a day ago	L5 TPM	15 yrs 6 yrs	$350,000 200K | 120K | 30K
Google Mountain View, CA | 2 days ago	L5 AI	5 yrs 1 yr	$247,000 180K | 40K | 27K
Google Mountain View, CA | 5 days ago	L5 General	25 yrs 6 yrs	$300,000 185K | 85K | 30K
Google Boston, MA | 12/07/2023	L5 TPM	12 yrs 2 yrs	$321,000 171K | 120K | 30K
Google New York, NY | 12/05/2023	L5 Project Management	18 yrs 2 yrs	$323,000 194K | 100K | 29K

Data Collection

Locate Needed Data

CHALLENGE: All the needed data is not located in one site
→ We have to crawl data from different endpoints

`/companies/:id` → Company Name, Company Size

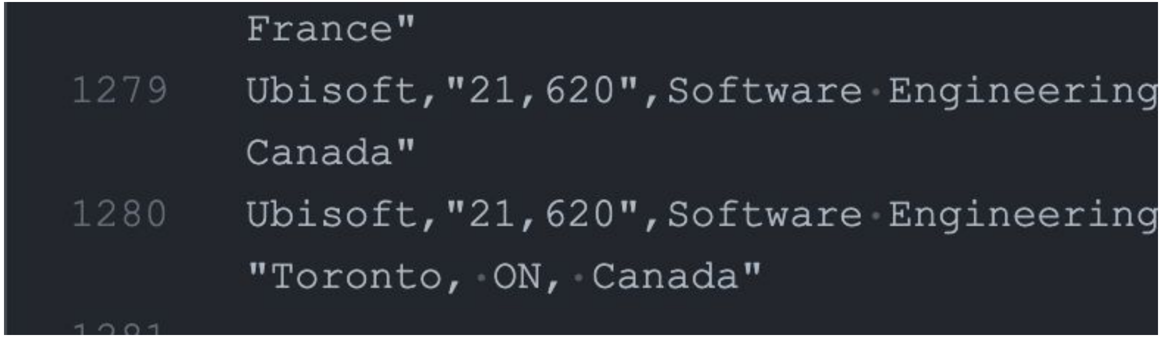
`/companies/:id/salaries` → Job Titles

`/companies/:id/salaries/:id` → Level, Domain, YOE, YOE at company, Base Salary, Stock, Bonus, Total Compensation, Location

Data Collection

Get Data

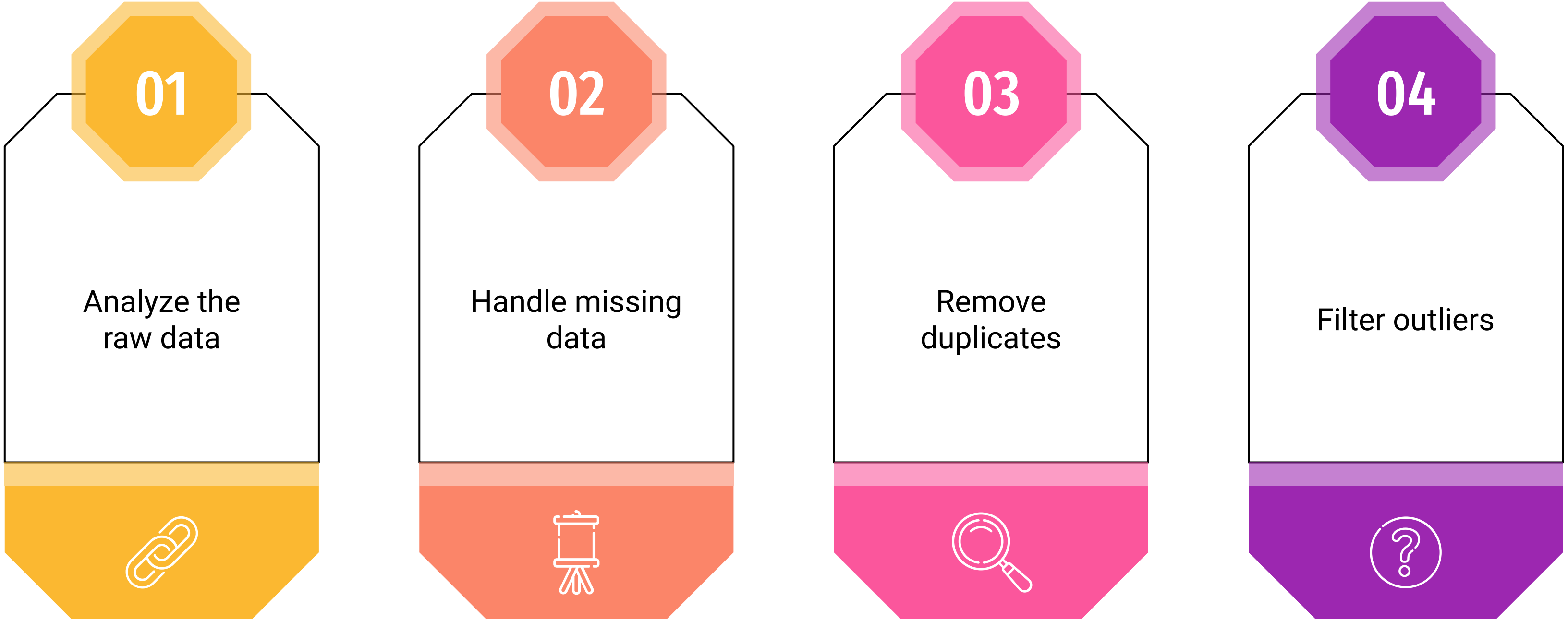
CHALLENGE: They have throttled the number of queries
→ We have to crawl data from different times



```
        France"
1279    Ubisoft,"21,620",Software·Engineering
        Canada"
1280    Ubisoft,"21,620",Software·Engineering
        "Toronto,·ON,·Canada"
```

1279 records were crawled within 2 days because levels.fyi throttled each IP, allowing only approximately 50 access attempts every 30 minutes.

Data Preprocessing



01

Analyze the raw data

02

Handle missing data

03

Remove duplicates

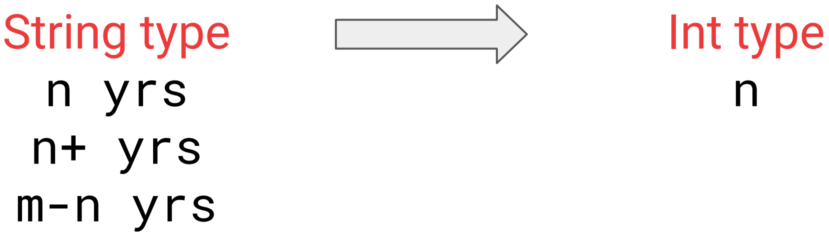
04

Filter outliers

Data Preprocessing

Due to the well-structured source data and a small dataset of just over 1,000 records, our preprocessing will primarily involve data cleaning, eliminating the need for complex procedures like data integration or transformation.

One of the main tasks is to standardize the YOE and YOE at company into a format:



String type

n yrs

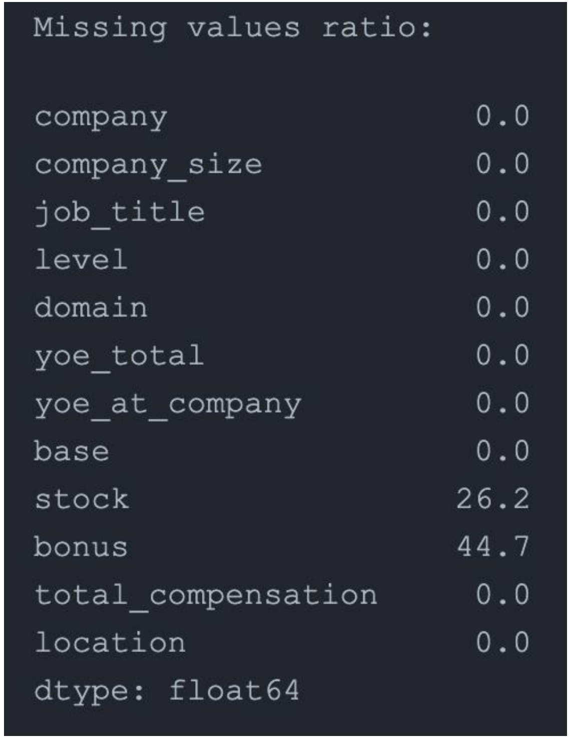
n+ yrs

m-n yrs

Int type

n

Data Preprocessing



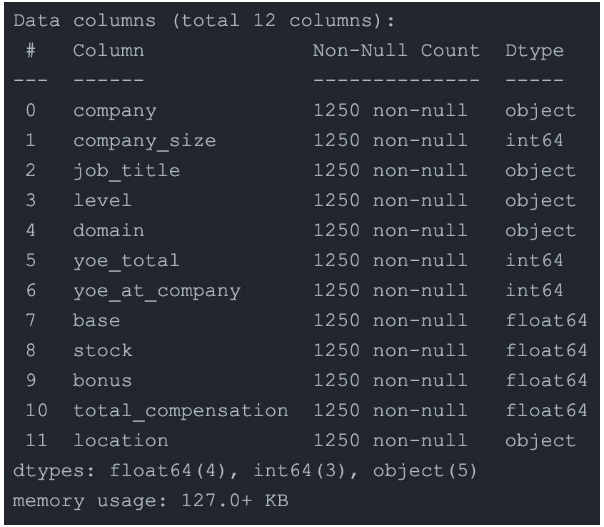
```
Missing values ratio:

company                0.0
company_size           0.0
job_title              0.0
level                  0.0
domain                 0.0
yoe_total              0.0
yoe_at_company         0.0
base                   0.0
stock                 26.2
bonus                 44.7
total_compensation     0.0
location               0.0
dtype: float64
```

With a missing ratio of 26.2% for the stock column and 44.7% for the bonus column, despite the very high missing data rates, these are not data that every company can provide publicly.

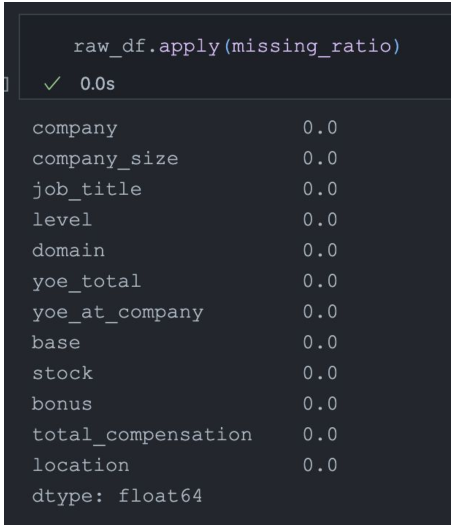
Moreover, they are used to validate the `base` and `total_compensation`, so we have decided to fill the missing values with 0.

Data Preprocessing

After performing additional checks such as ensuring the validity of `total_compensation` and converting numerical data to integers and floats, we obtain the pre-processed data as follows.



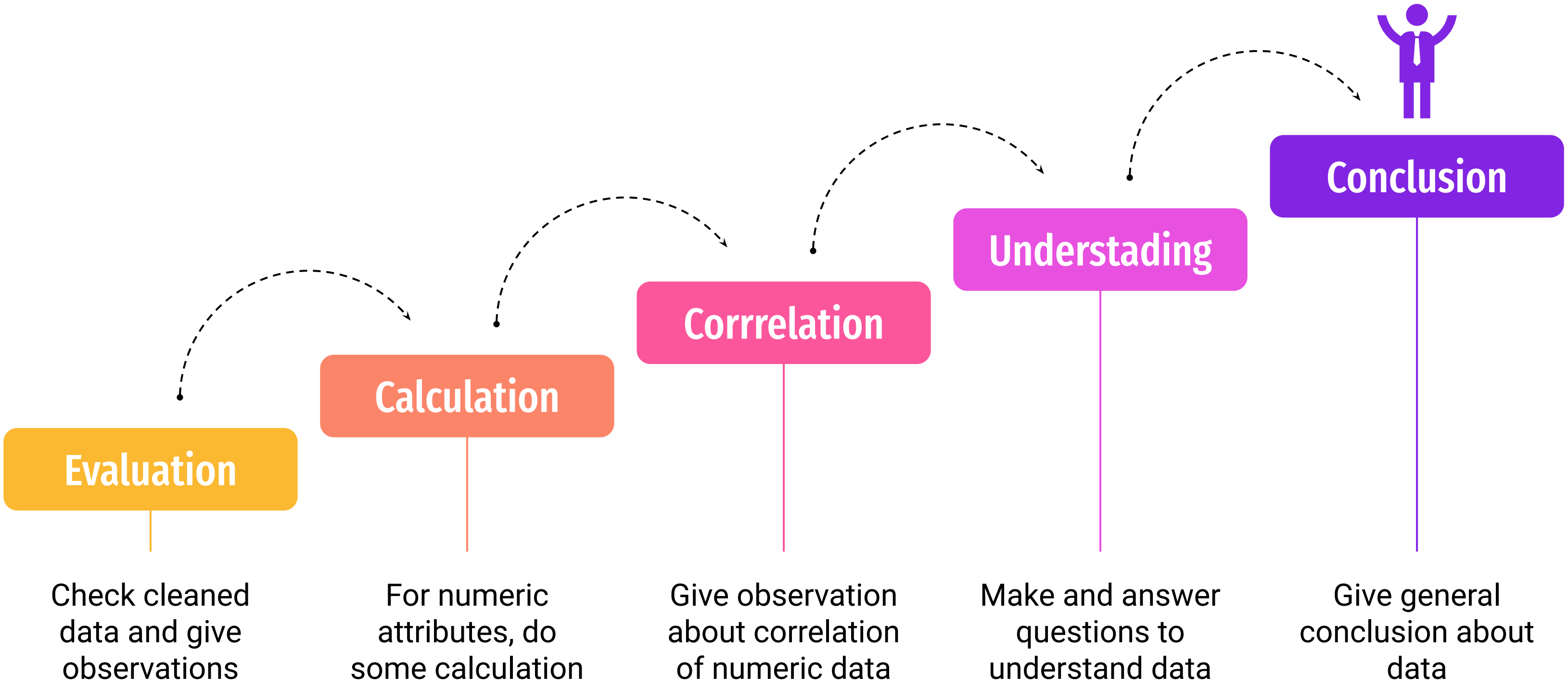
```
Data columns (total 12 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   company             1250 non-null   object
 1   company_size        1250 non-null   int64
 2   job_title           1250 non-null   object
 3   level               1250 non-null   object
 4   domain              1250 non-null   object
 5   yoe_total           1250 non-null   int64
 6   yoe_at_company      1250 non-null   int64
 7   base                1250 non-null   float64
 8   stock               1250 non-null   float64
 9   bonus               1250 non-null   float64
 10  total_compensation  1250 non-null   float64
 11  location            1250 non-null   object
dtypes: float64(4), int64(3), object(5)
memory usage: 127.0+ KB
```



```
raw_df.apply(missing_ratio)
```

```
company               0.0
company_size          0.0
job_title             0.0
level                 0.0
domain                0.0
yoe_total             0.0
yoe_at_company        0.0
base                  0.0
stock                 0.0
bonus                 0.0
total_compensation    0.0
location              0.0
dtype: float64
```

Data Exploration



Evaluation

Check cleaned data and give observations

Calculation

For numeric attributes, do some calculation

Corrrelation

Give observation about correlation of numeric data

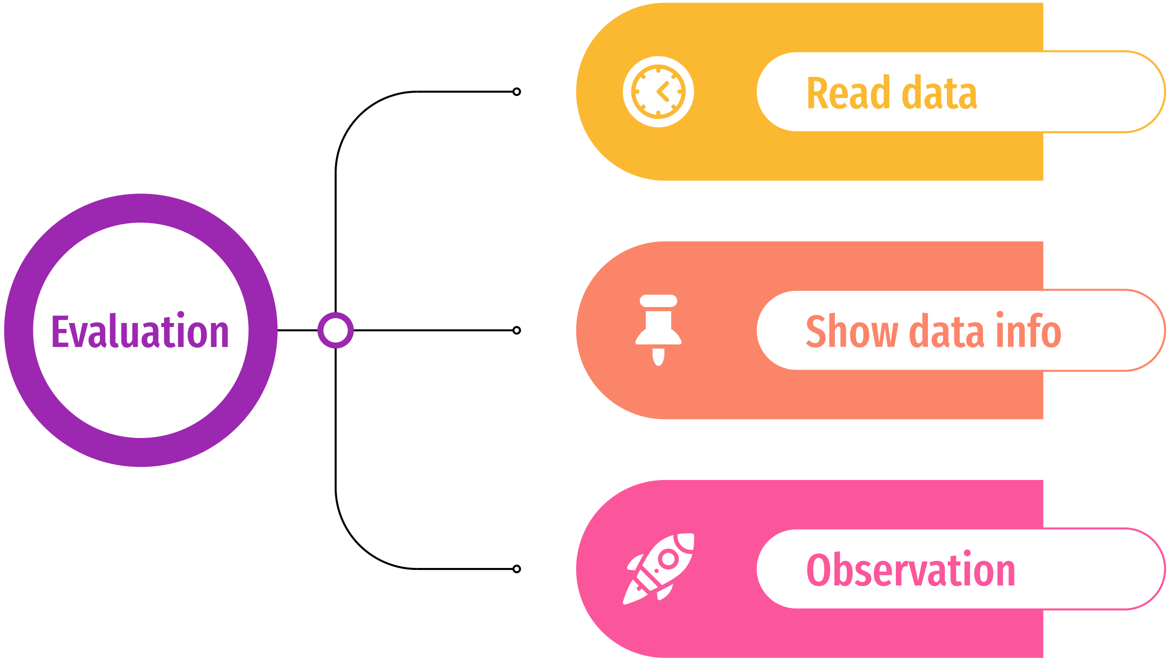
Understading

Make and answer questions to understand data

Conclusion

Give general conclusion about data

Evaluation



Evaluation

Read data

Read data from file .csv and print data

Show data info

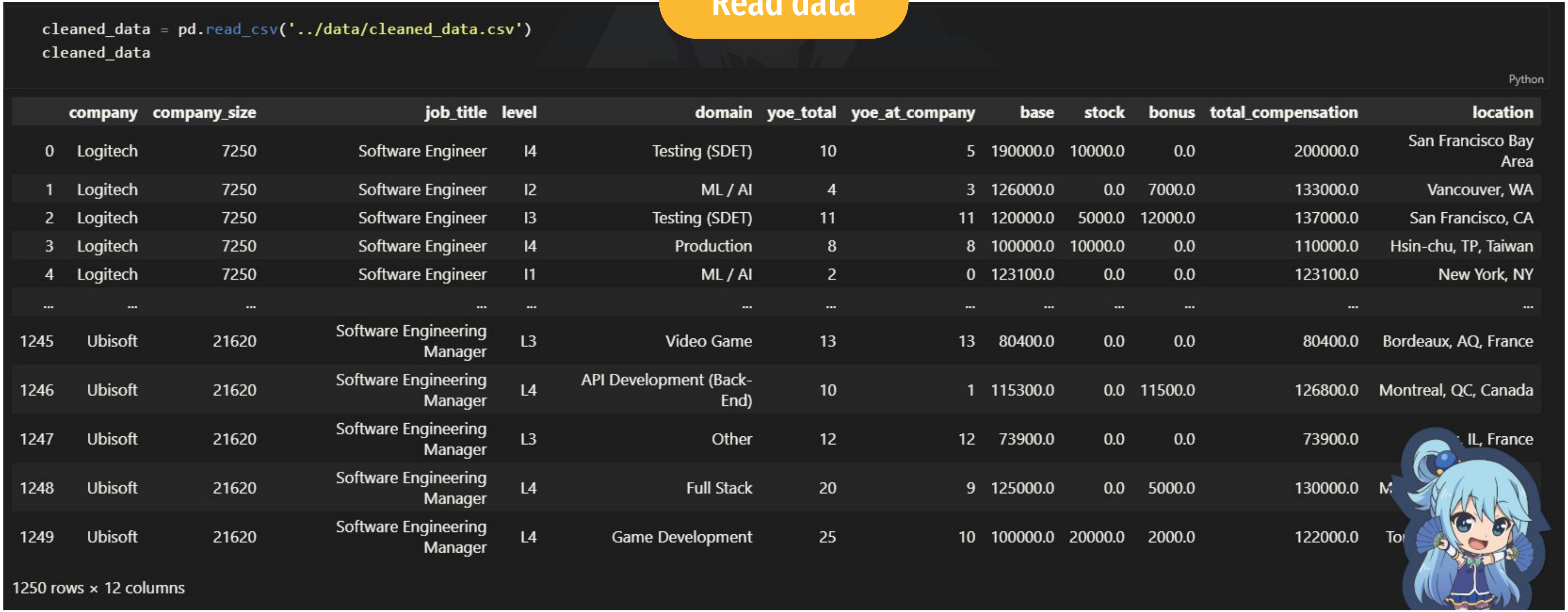
Print data info: name of attributes, non-null count and dtypes

Observation

Give observation

Evaluation

Read data



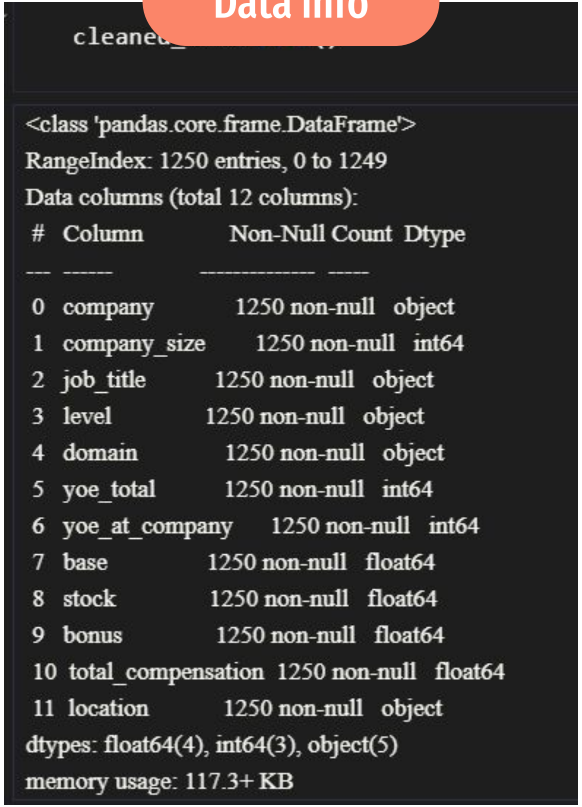
```python
cleaned_data = pd.read_csv('../data/cleaned_data.csv')
cleaned_data
```

	company	company_size	job_title	level	domain	yoe_total	yoe_at_company	base	stock	bonus	total_compensation	location
0	Logitech	7250	Software Engineer	I4	Testing (SDET)	10	5	190000.0	10000.0	0.0	200000.0	San Francisco Bay Area
1	Logitech	7250	Software Engineer	I2	ML / AI	4	3	126000.0	0.0	7000.0	133000.0	Vancouver, WA
2	Logitech	7250	Software Engineer	I3	Testing (SDET)	11	11	120000.0	5000.0	12000.0	137000.0	San Francisco, CA
3	Logitech	7250	Software Engineer	I4	Production	8	8	100000.0	10000.0	0.0	110000.0	Hsin-chu, TP, Taiwan
4	Logitech	7250	Software Engineer	I1	ML / AI	2	0	123100.0	0.0	0.0	123100.0	New York, NY
...	...	...	...	...	...	...	...	...	...	...	...	...
1245	Ubisoft	21620	Software Engineering Manager	L3	Video Game	13	13	80400.0	0.0	0.0	80400.0	Bordeaux, AQ, France
1246	Ubisoft	21620	Software Engineering Manager	L4	API Development (Back-End)	10	1	115300.0	0.0	11500.0	126800.0	Montreal, QC, Canada
1247	Ubisoft	21620	Software Engineering Manager	L3	Other	12	12	73900.0	0.0	0.0	73900.0	IL, France
1248	Ubisoft	21620	Software Engineering Manager	L4	Full Stack	20	9	125000.0	0.0	5000.0	130000.0	M
1249	Ubisoft	21620	Software Engineering Manager	L4	Game Development	25	10	100000.0	20000.0	2000.0	122000.0	To

1250 rows × 12 columns

Evaluation

Data info

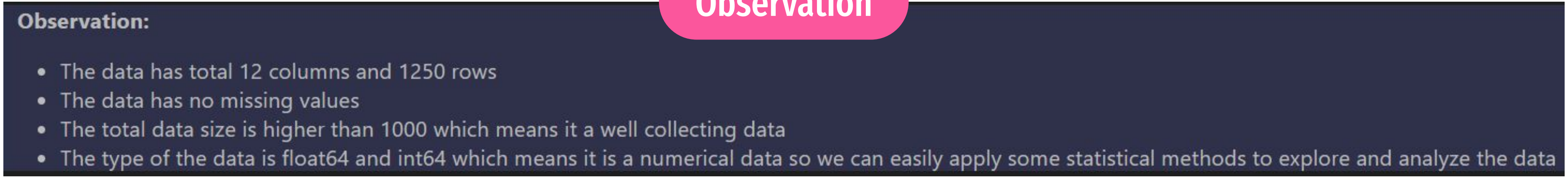


```
cleaned_
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1250 entries, 0 to 1249
Data columns (total 12 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   company             1250 non-null   object
 1   company_size        1250 non-null   int64
 2   job_title           1250 non-null   object
 3   level               1250 non-null   object
 4   domain              1250 non-null   object
 5   yoe_total           1250 non-null   int64
 6   yoe_at_company      1250 non-null   int64
 7   base                1250 non-null   float64
 8   stock               1250 non-null   float64
 9   bonus               1250 non-null   float64
 10  total_compensation  1250 non-null   float64
 11  location            1250 non-null   object
dtypes: float64(4), int64(3), object(5)
memory usage: 117.3+ KB
```

Evaluation

Observation



Observation:

- The data has total 12 columns and 1250 rows
- The data has no missing values
- The total data size is higher than 1000 which means it a well collecting data
- The type of the data is float64 and int64 which means it is a numerical data so we can easily apply some statistical methods to explore and analyze the data

Calculation

Calculate numeric data

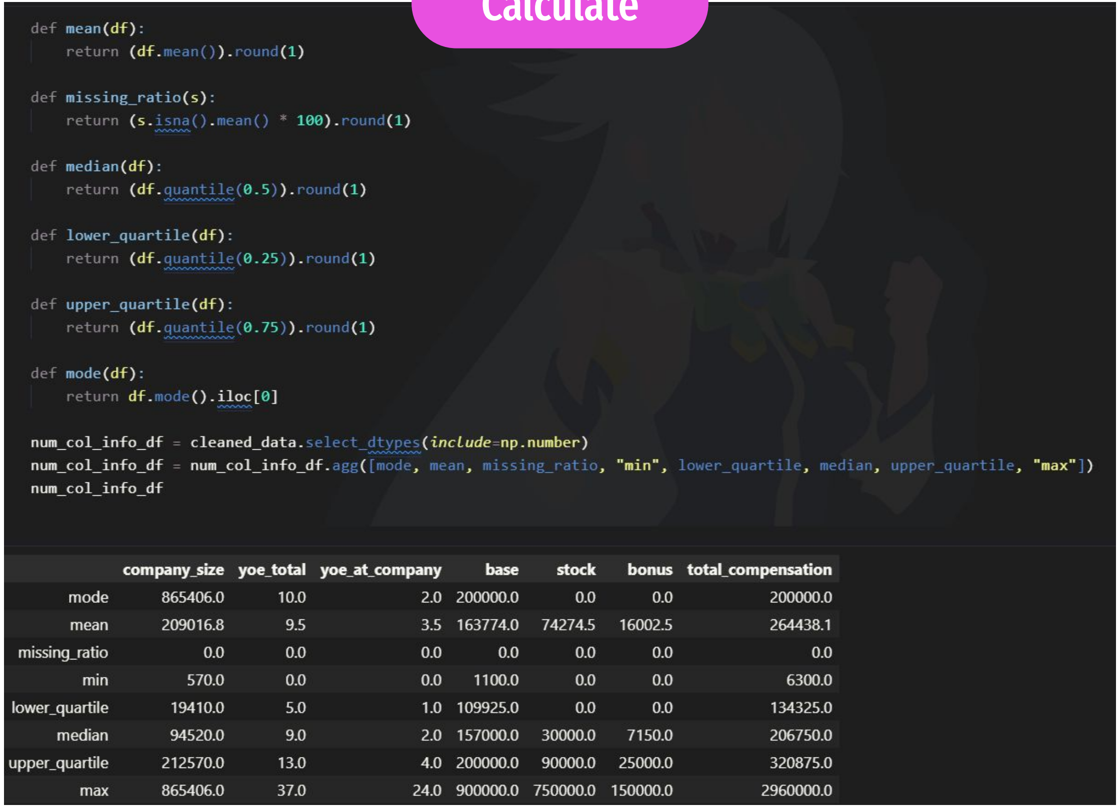
Calculate mean, median, lower quartile, upper quartile and mode of numeric data

Give observation

Give observation about mean, median, mode, max, min

Calculation

Calculate

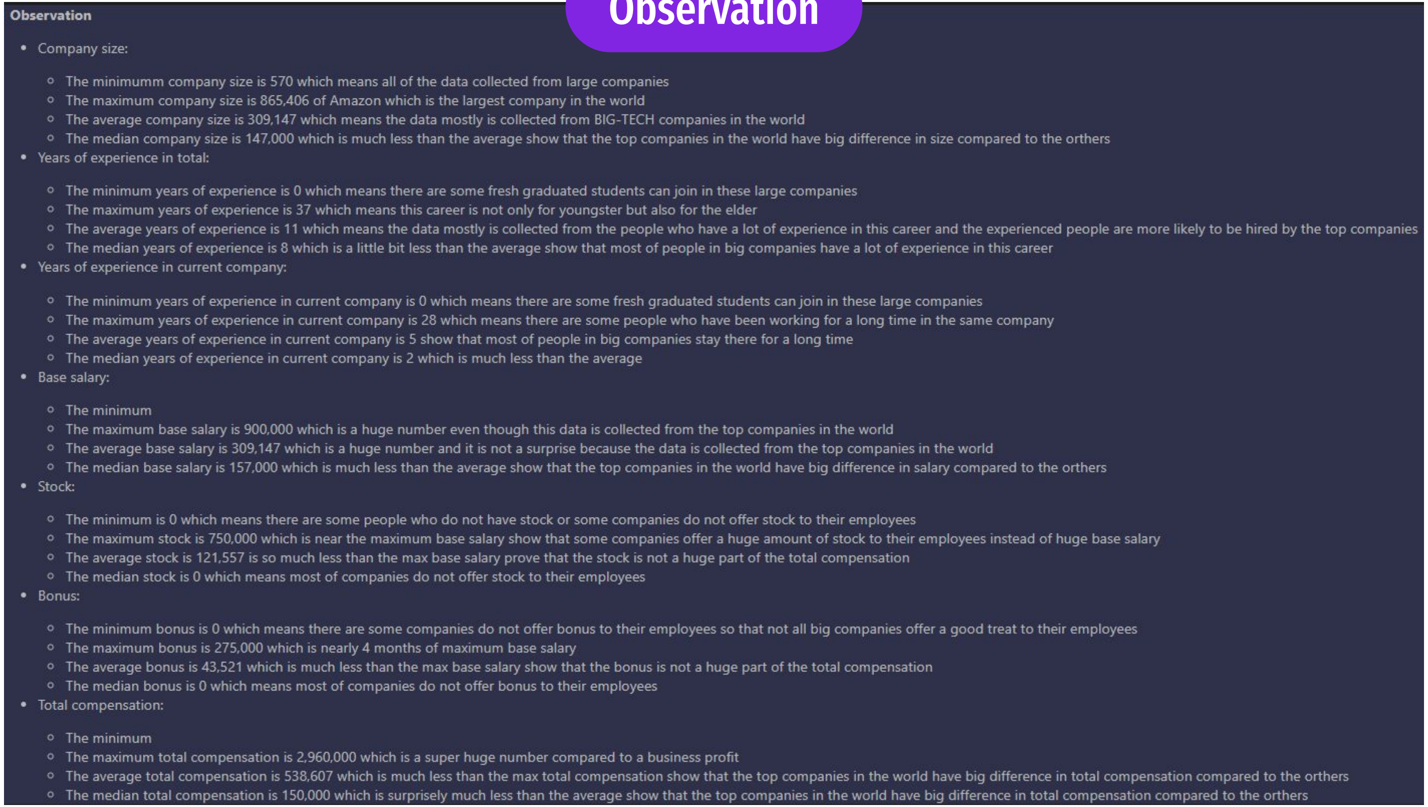


```
def mean(df):
    return (df.mean()).round(1)

def missing_ratio(s):
    return (s.isna().mean() * 100).round(1)

def median(df):
    return (df.quantile(0.5)).round(1)

def lower_quartile(df):
    return (df.quantile(0.25)).round(1)

def upper_quartile(df):
    return (df.quantile(0.75)).round(1)

def mode(df):
    return df.mode().iloc[0]

num_col_info_df = cleaned_data.select_dtypes(include=np.number)
num_col_info_df = num_col_info_df.agg([mode, mean, missing_ratio, "min", lower_quartile, median, upper_quartile, "max"])
num_col_info_df
```

	company_size	yoe_total	yoe_at_company	base	stock	bonus	total_compensation
mode	865406.0	10.0	2.0	200000.0	0.0	0.0	200000.0
mean	209016.8	9.5	3.5	163774.0	74274.5	16002.5	264438.1
missing_ratio	0.0	0.0	0.0	0.0	0.0	0.0	0.0
min	570.0	0.0	0.0	1100.0	0.0	0.0	6300.0
lower_quartile	19410.0	5.0	1.0	109925.0	0.0	0.0	134325.0
median	94520.0	9.0	2.0	157000.0	30000.0	7150.0	206750.0
upper_quartile	212570.0	13.0	4.0	200000.0	90000.0	25000.0	320875.0
max	865406.0	37.0	24.0	900000.0	750000.0	150000.0	2960000.0

Calculation

Observation



Observation

- Company size:
    - The minimumm company size is 570 which means all of the data collected from large companies
    - The maximum company size is 865,406 of Amazon which is the largest company in the world
    - The average company size is 309,147 which means the data mostly is collected from BIG-TECH companies in the world
    - The median company size is 147,000 which is much less than the average show that the top companies in the world have big difference in size compared to the orthers
- Years of experience in total:
    - The minimum years of experience is 0 which means there are some fresh graduated students can join in these large companies
    - The maximum years of experience is 37 which means this career is not only for youngster but also for the elder
    - The average years of experience is 11 which means the data mostly is collected from the people who have a lot of experience in this career and the experienced people are more likely to be hired by the top companies
    - The median years of experience is 8 which is a little bit less than the average show that most of people in big companies have a lot of experience in this career
- Years of experience in current company:
    - The minimum years of experience in current company is 0 which means there are some fresh graduated students can join in these large companies
    - The maximum years of experience in current company is 28 which means there are some people who have been working for a long time in the same company
    - The average years of experience in current company is 5 show that most of people in big companies stay there for a long time
    - The median years of experience in current company is 2 which is much less than the average
- Base salary:
    - The minimum
    - The maximum base salary is 900,000 which is a huge number even though this data is collected from the top companies in the world
    - The average base salary is 309,147 which is a huge number and it is not a surprise because the data is collected from the top companies in the world
    - The median base salary is 157,000 which is much less than the average show that the top companies in the world have big difference in salary compared to the orthers
- Stock:
    - The minimum is 0 which means there are some people who do not have stock or some companies do not offer stock to their employees
    - The maximum stock is 750,000 which is near the maximum base salary show that some companies offer a huge amount of stock to their employees instead of huge base salary
    - The average stock is 121,557 is so much less than the max base salary prove that the stock is not a huge part of the total compensation
    - The median stock is 0 which means most of companies do not offer stock to their employees
- Bonus:
    - The minimum bonus is 0 which means there are some companies do not offer bonus to their employees so that not all big companies offer a good treat to their employees
    - The maximum bonus is 275,000 which is nearly 4 months of maximum base salary
    - The average bonus is 43,521 which is much less than the max base salary show that the bonus is not a huge part of the total compensation
    - The median bonus is 0 which means most of companies do not offer bonus to their employees
- Total compensation:
    - The minimum
    - The maximum total compensation is 2,960,000 which is a super huge number compared to a business profit
    - The average total compensation is 538,607 which is much less than the max total compensation show that the top companies in the world have big difference in total compensation compared to the orthers
    - The median total compensation is 150,000 which is surprisely much less than the average show that the top companies in the world have big difference in total compensation compared to the orthers

Correlation

Show heatmap

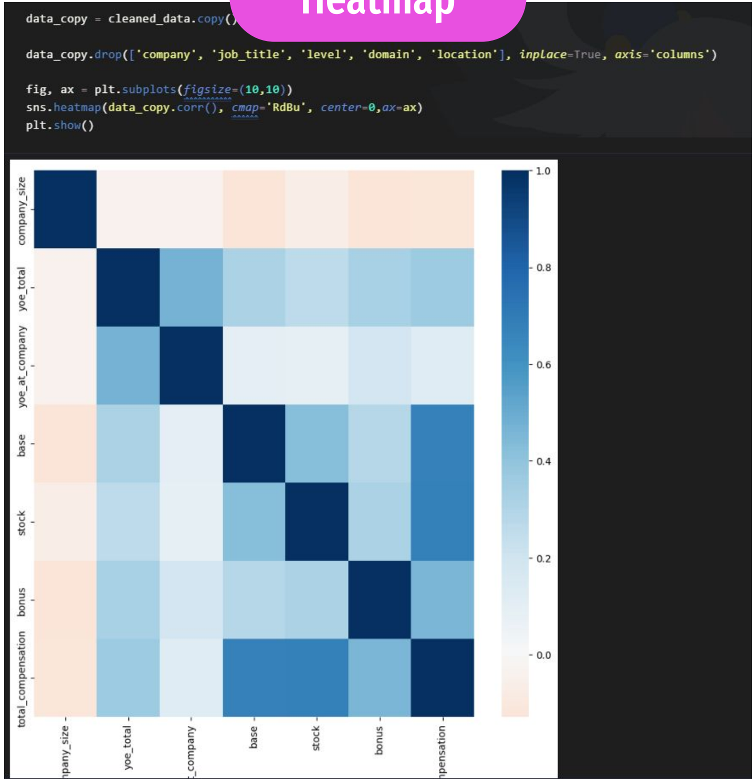
Create heatmap using numeric data to show correlation

Give observation

Give observation about correlation of data attributes

Calculation

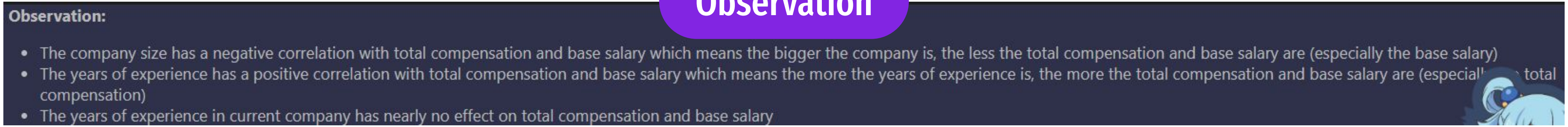
Heatmap



```
data_copy = cleaned_data.copy()

data_copy.drop(['company', 'job_title', 'level', 'domain', 'location'], inplace=True, axis='columns')

fig, ax = plt.subplots(figsize=(10,10))
sns.heatmap(data_copy.corr(), cmap='RdBu', center=0,ax=ax)
plt.show()
```

Calculation

Observation



Observation:

- The company size has a negative correlation with total compensation and base salary which means the bigger the company is, the less the total compensation and base salary are (especially the base salary)
- The years of experience has a positive correlation with total compensation and base salary which means the more the years of experience is, the more the total compensation and base salary are (especiall total compensation)
- The years of experience in current company has nearly no effect on total compensation and base salary

Understanding

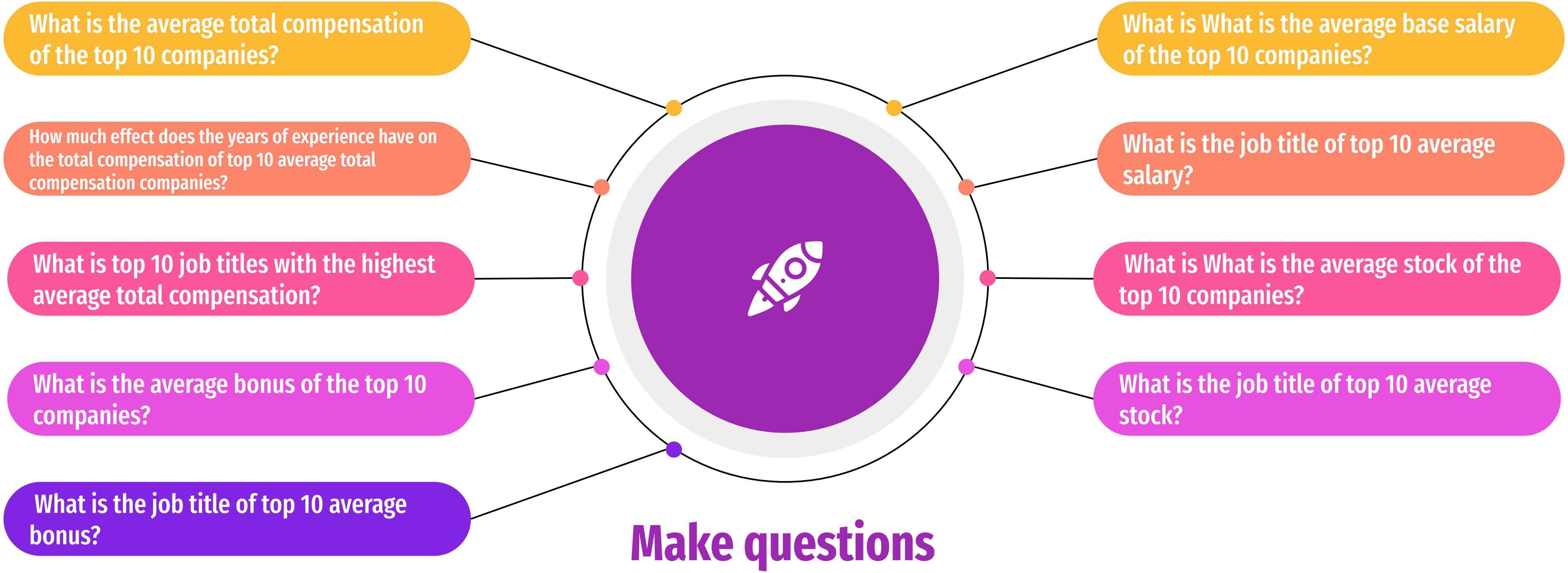
Make questions

Make some questions about data

Answer questions

Answer questions to understand data

Understanding



What is the average total compensation of the top 10 companies?

How much effect does the years of experience have on the total compensation of top 10 average total compensation companies?

What is top 10 job titles with the highest average total compensation?

What is the average bonus of the top 10 companies?

What is the job title of top 10 average bonus?

What is What is the average base salary of the top 10 companies?

What is the job title of top 10 average salary?

What is What is the average stock of the top 10 companies?

What is the job title of top 10 average stock?

Make questions

Understanding

Step 1

Define

Define the question (if needed)

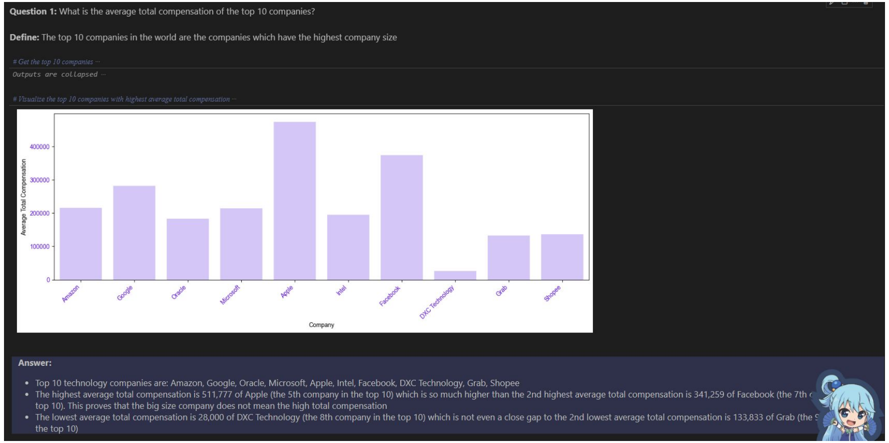
Step 2

Answer

Answer the question (visualize the answer)

Answer questions

Understanding



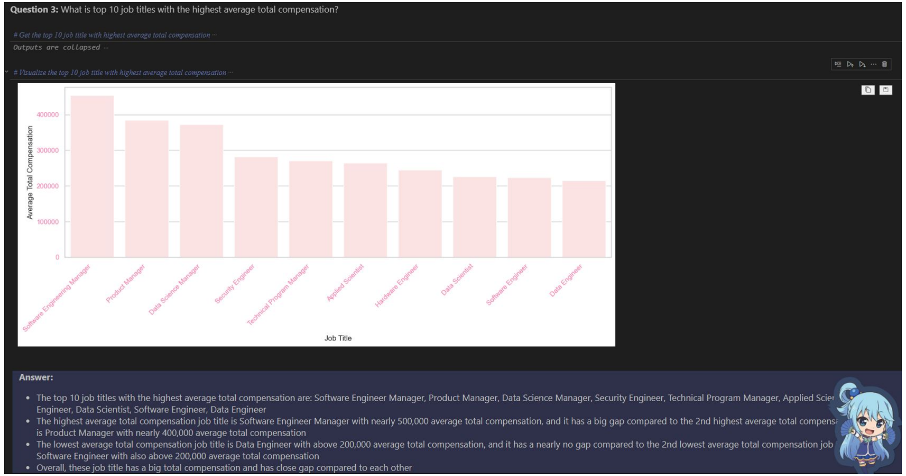
Question 1: What is the average total compensation of the top 10 companies?

Define: The top 10 companies in the world are the companies which have the highest company size

```
# Get the top 10 companies ···
Outputs are collapsed ···

# Visualize the top 10 companies with highest average total compensation ···
```

Answer:

- Top 10 technology companies are: Amazon, Google, Oracle, Microsoft, Apple, Intel, Facebook, DXC Technology, Grab, Shopee
- The highest average total compensation is 511,777 of Apple (the 5th company in the top 10) which is so much higher than the 2nd highest average total compensation is 341,259 of Facebook (the 7th c top 10). This proves that the big size company does not mean the high total compensation
- The lowest average total compensation is 28,000 of DXC Technology (the 8th company in the top 10) which is not even a close gap to the 2nd lowest average total compensation is 133,833 of Grab (the 9 the top 10)



Question 3: What is top 10 job titles with the highest average total compensation?

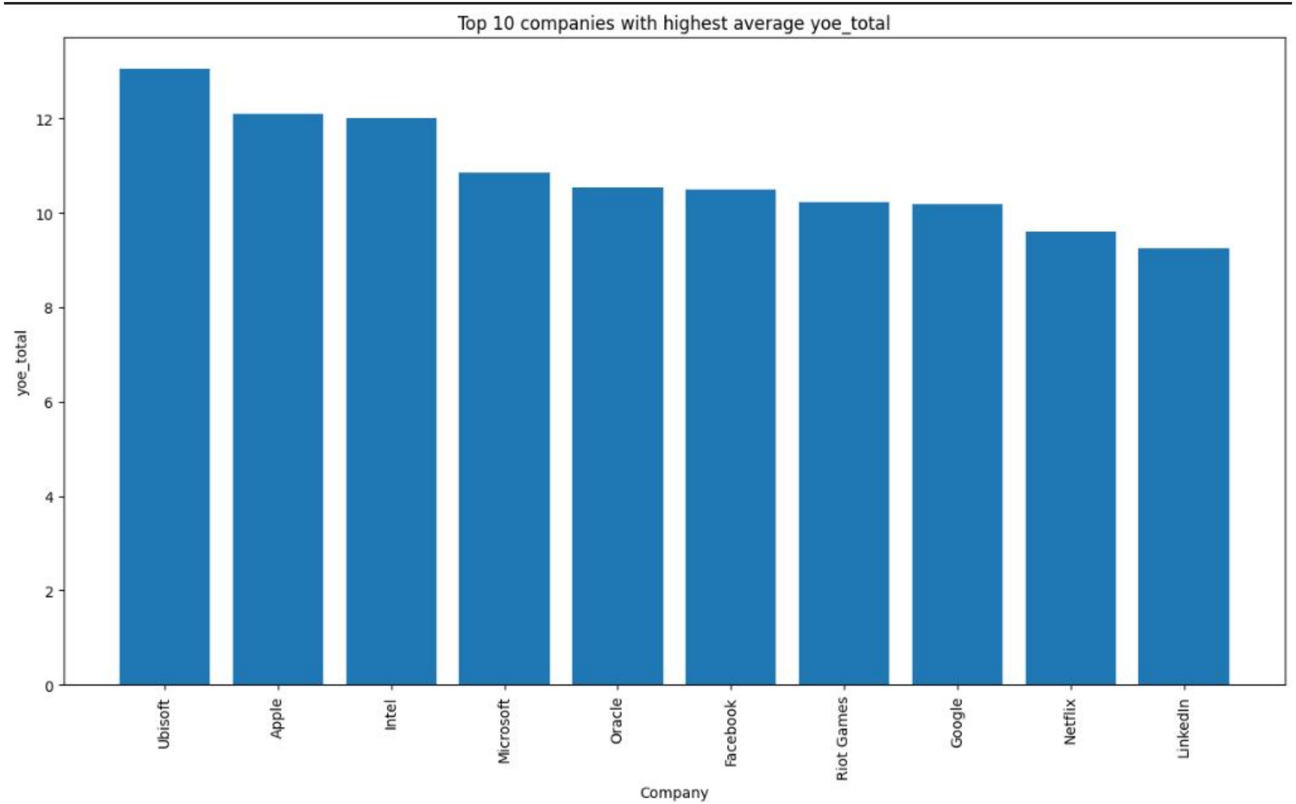
```
# Get the top 10 job title with highest average total compensation ···
Outputs are collapsed ···

# Visualize the top 10 job title with highest average total compensation ···
```

Answer:

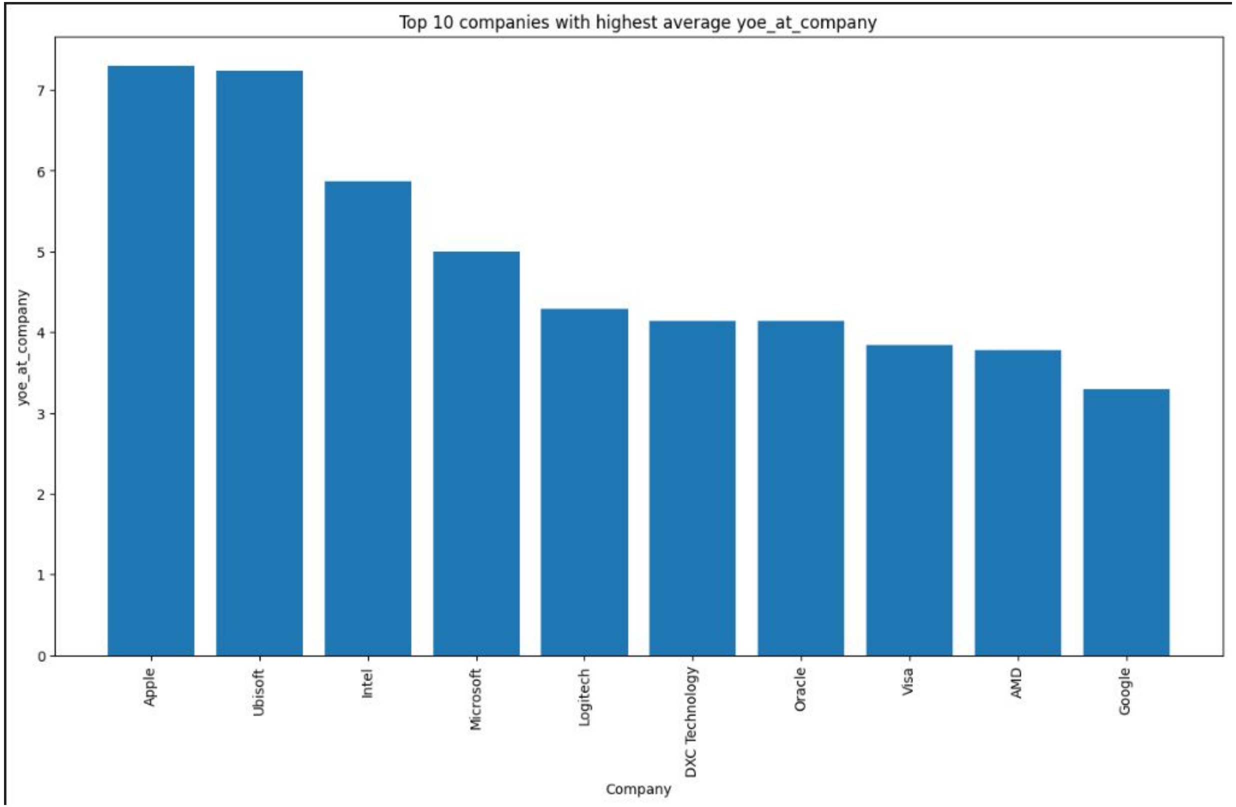
- The top 10 job titles with the highest average total compensation are: Software Engineer Manager, Product Manager, Data Science Manager, Security Engineer, Technical Program Manager, Applied Scie Engineer, Data Scientist, Software Engineer, Data Engineer
- The highest average total compensation job title is Software Engineer Manager with nearly 500,000 average total compensation, and it has a big gap compared to the 2nd highest average total compens is Product Manager with nearly 400,000 average total compensation
- The lowest average total compensation job title is Data Engineer with above 200,000 average total compensation, and it has a nearly no gap compared to the 2nd lowest average total compensation job Software Engineer with also above 200,000 average total compensation
- Overall, these job title has a big total compensation and has close gap compared to each other

Answer questions

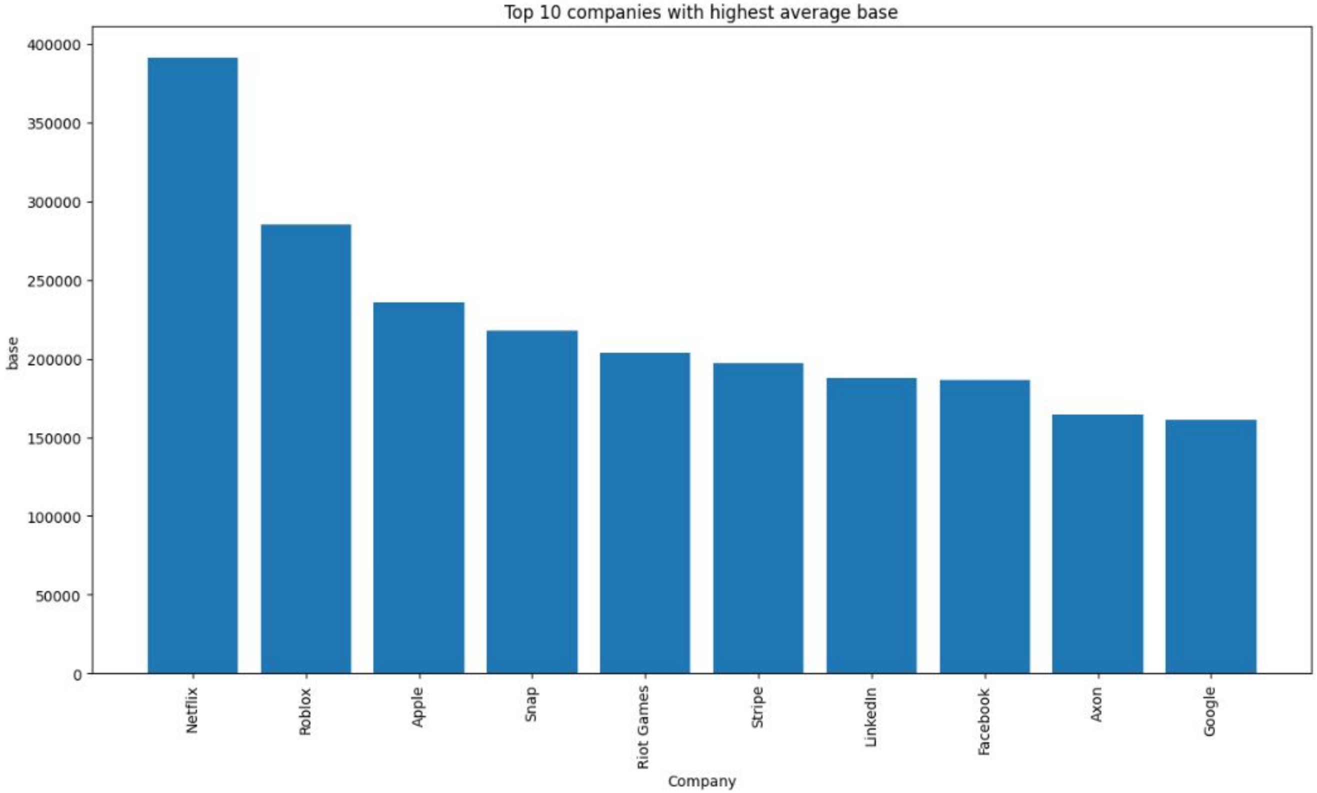
Visualization



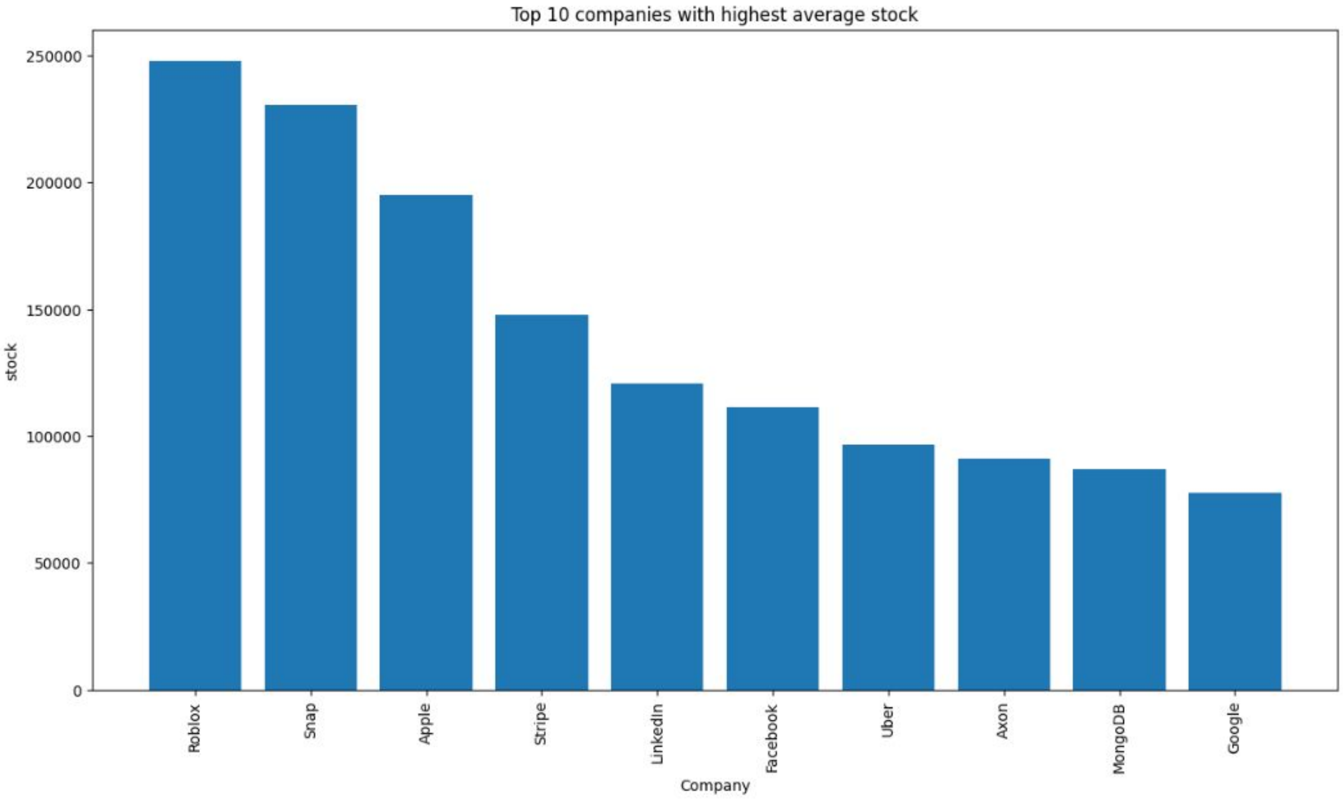
Visualization



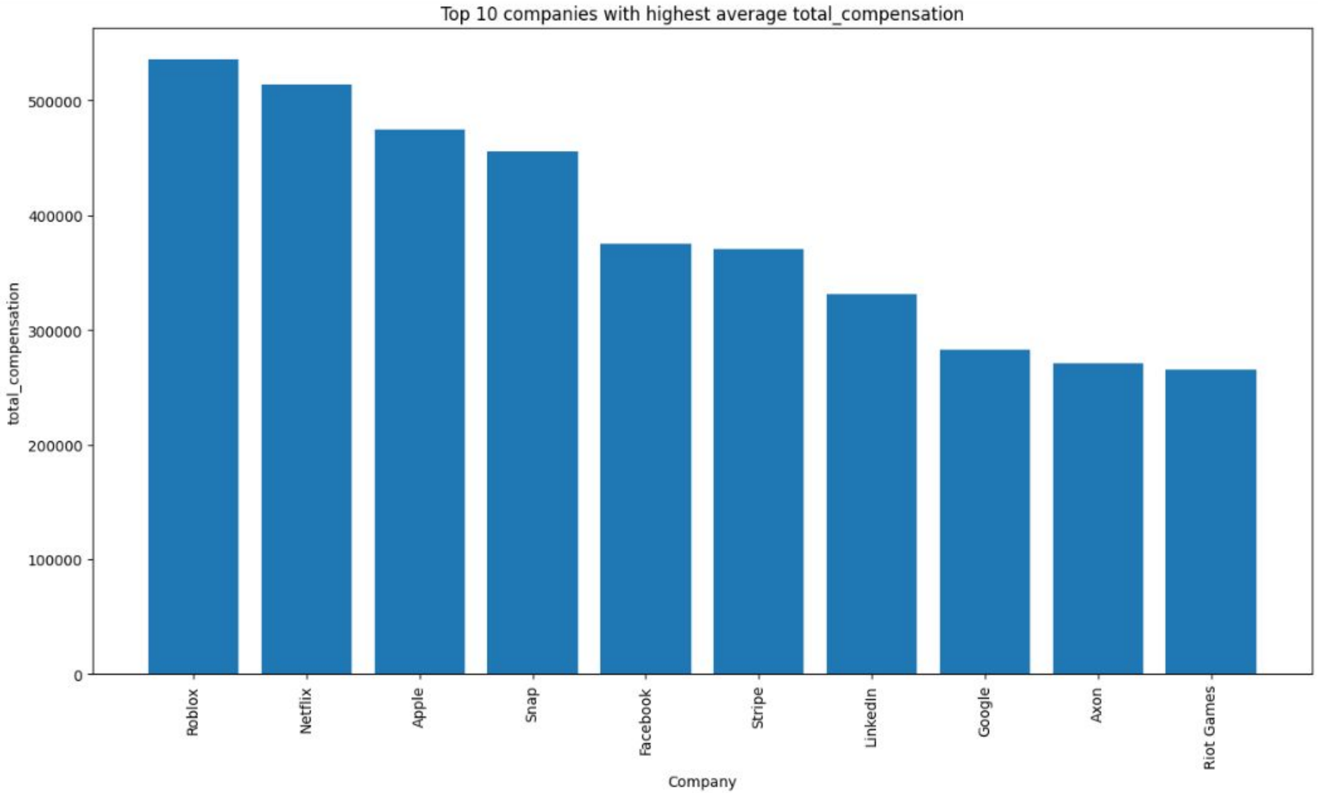
Visualization



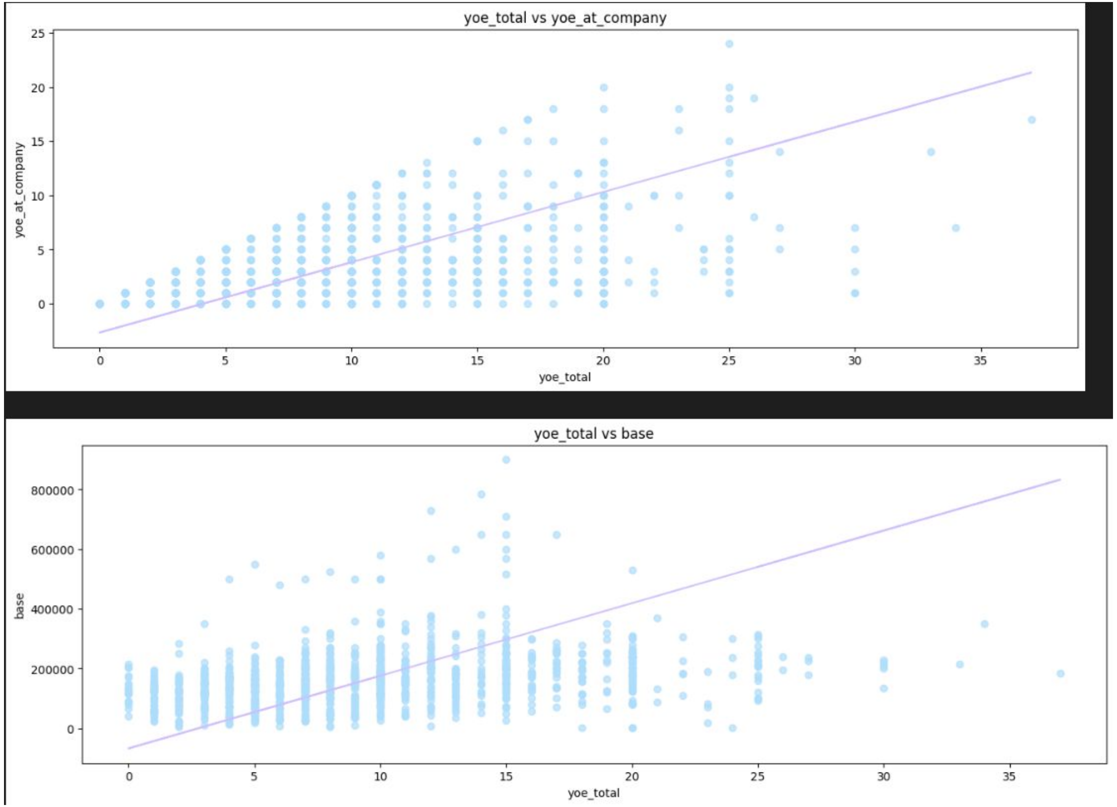
Visualization



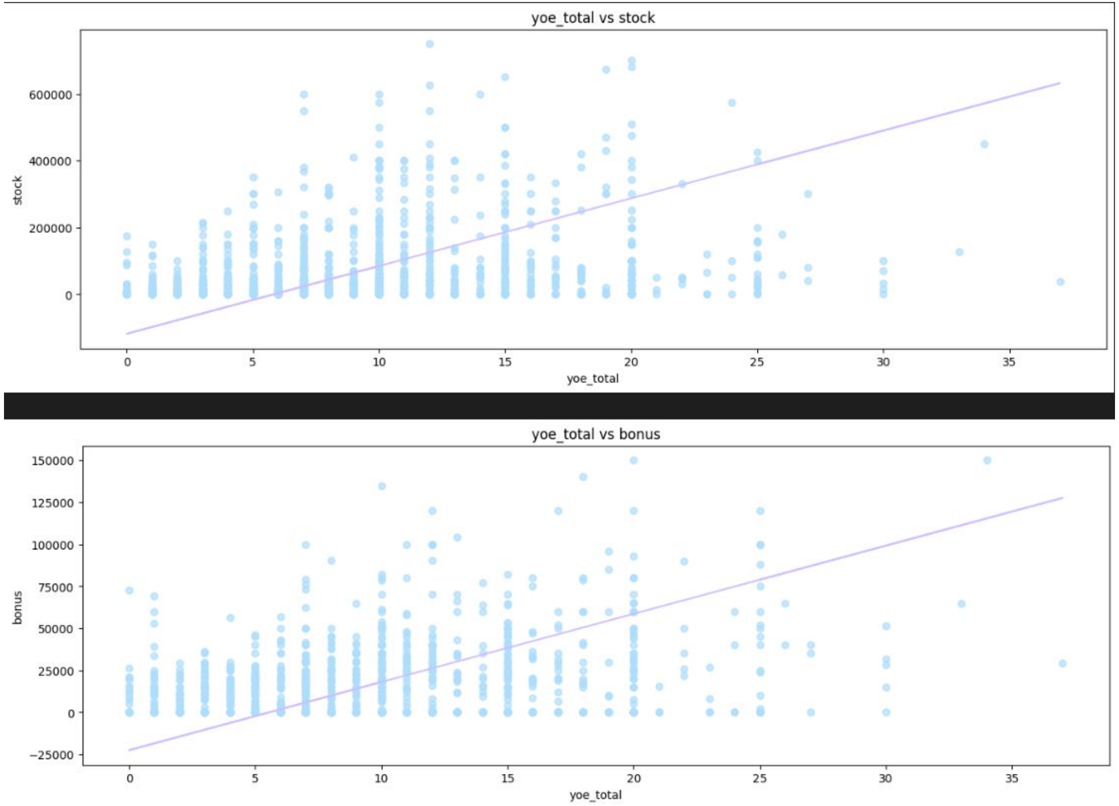
Visualization



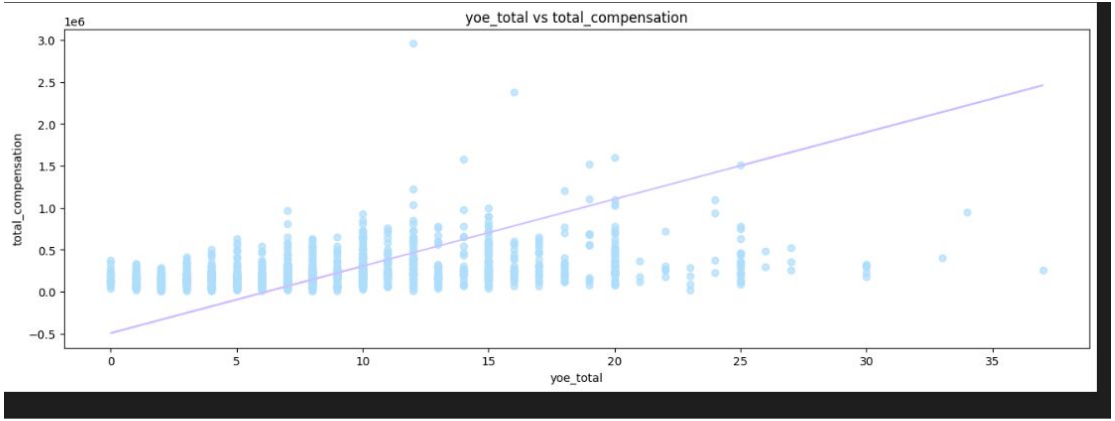
Visualization

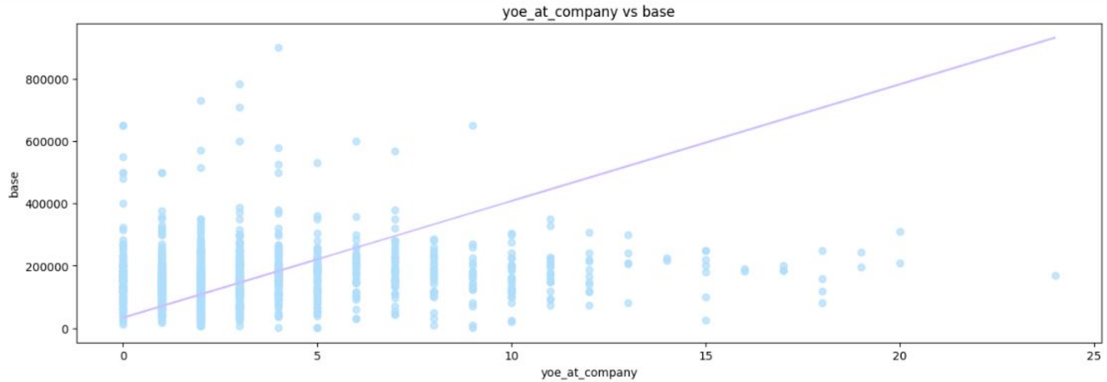


Visualization

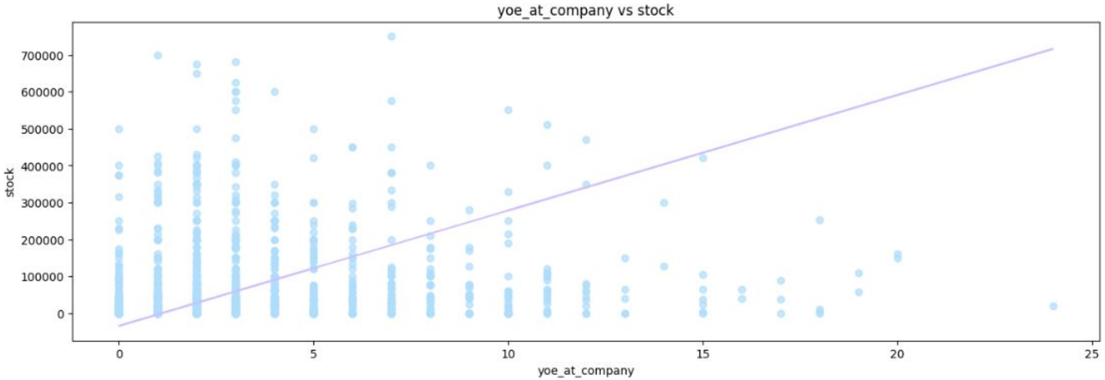


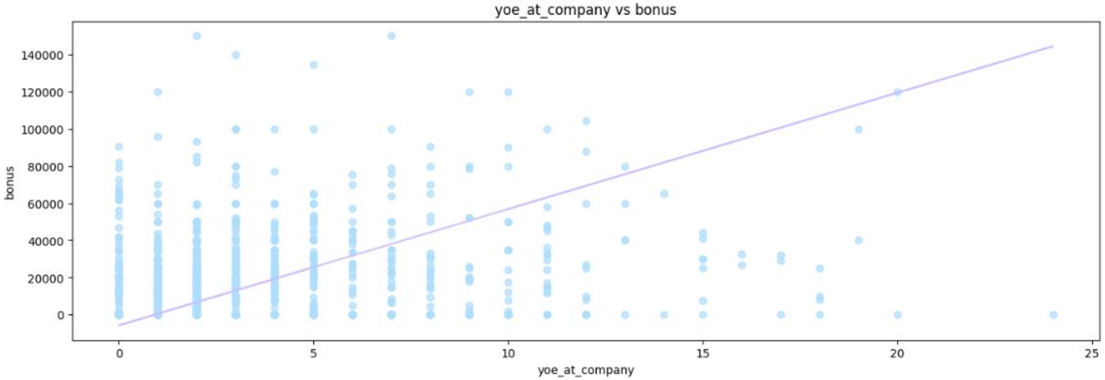
Visualization



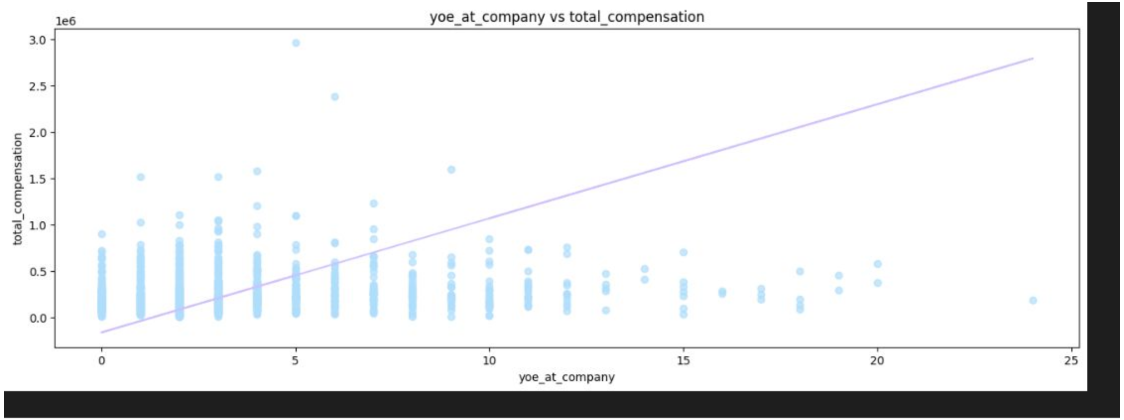


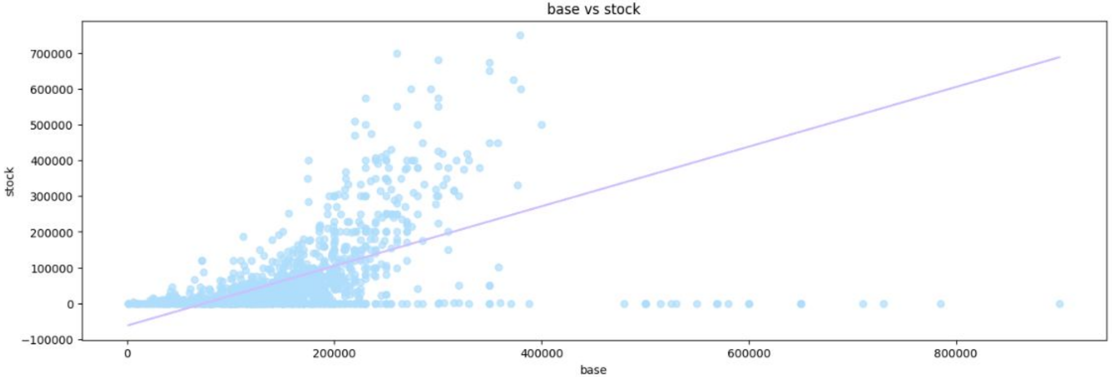
Visualization



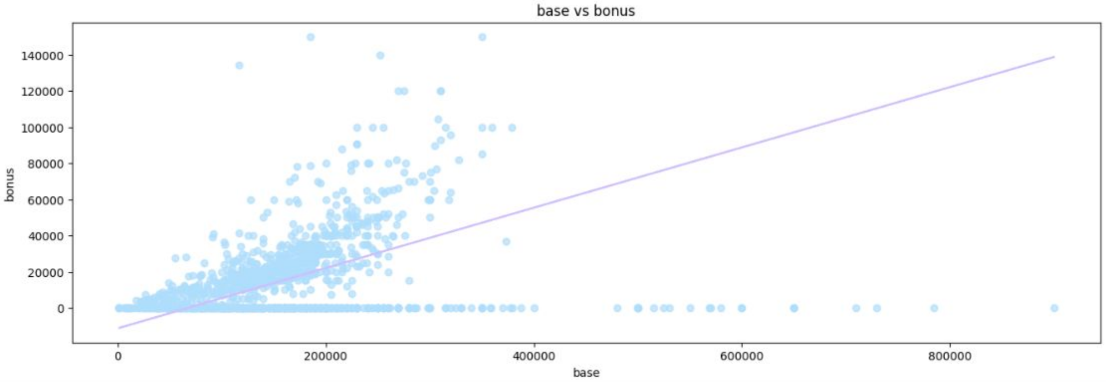


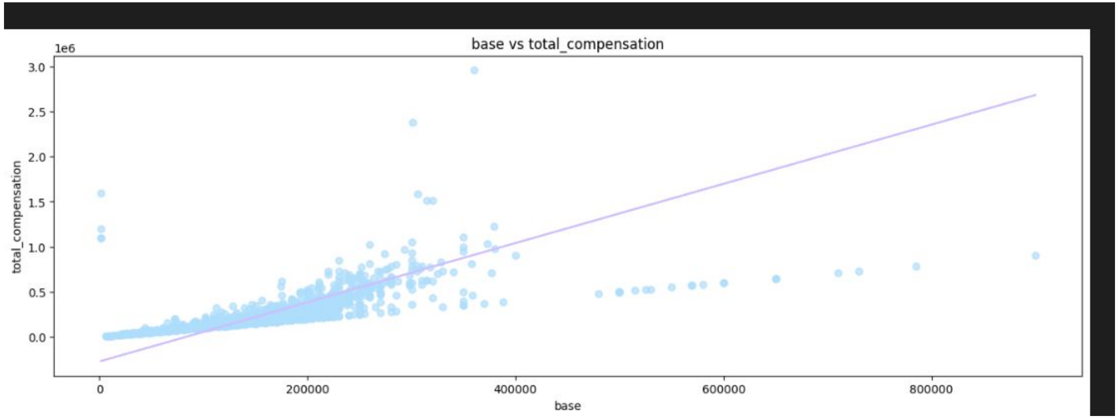
Visualization



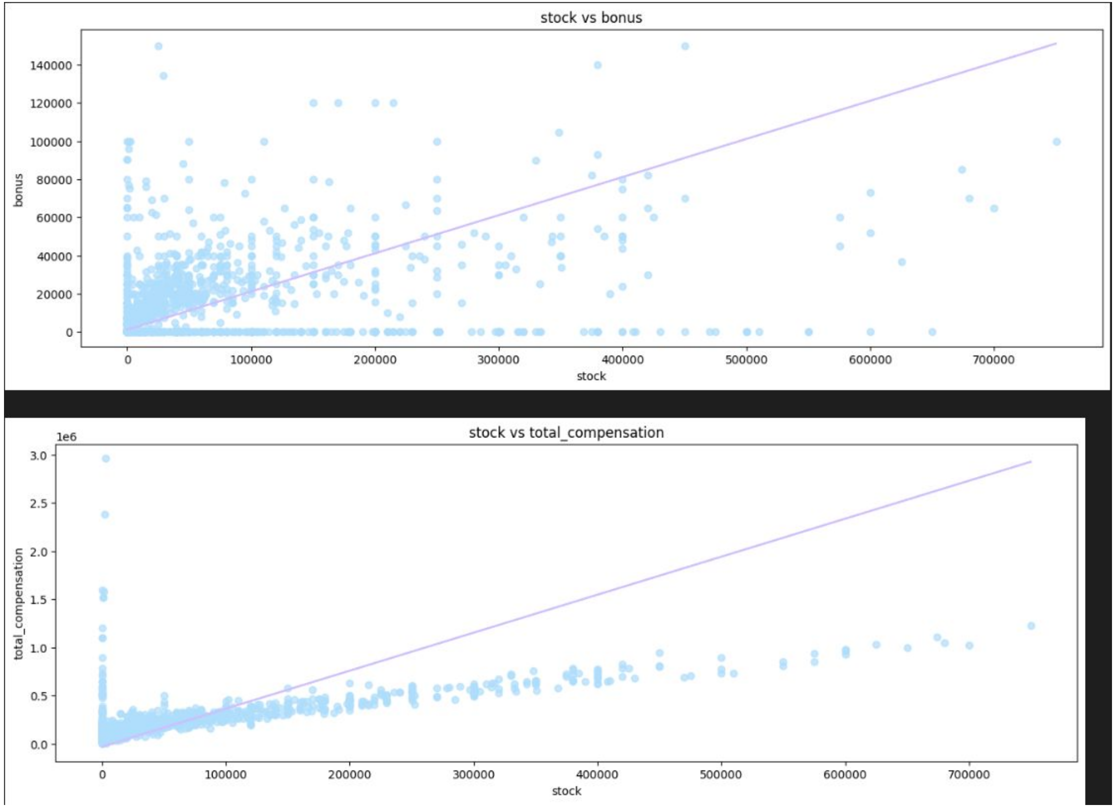


Visualization

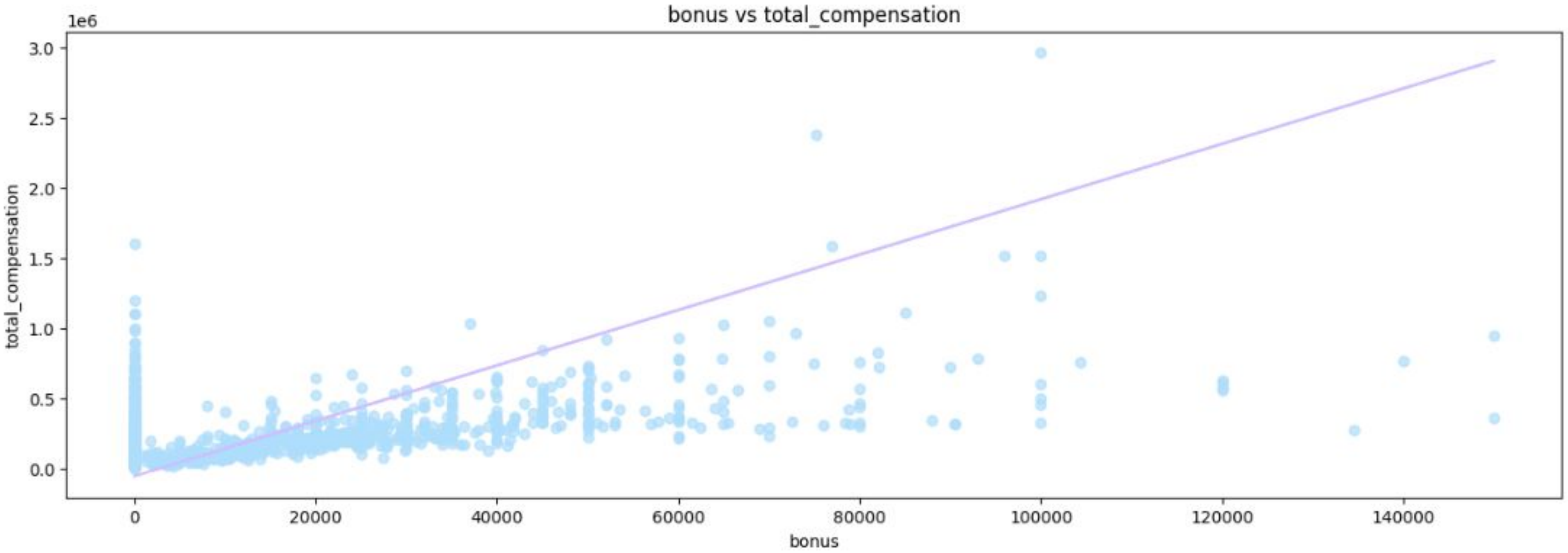




Visualization



Visualization



Data Modeling

Define problem

Define what problem to be solved

Prepare Data

Prepare data for training and testing

Visualization

Visualize the result

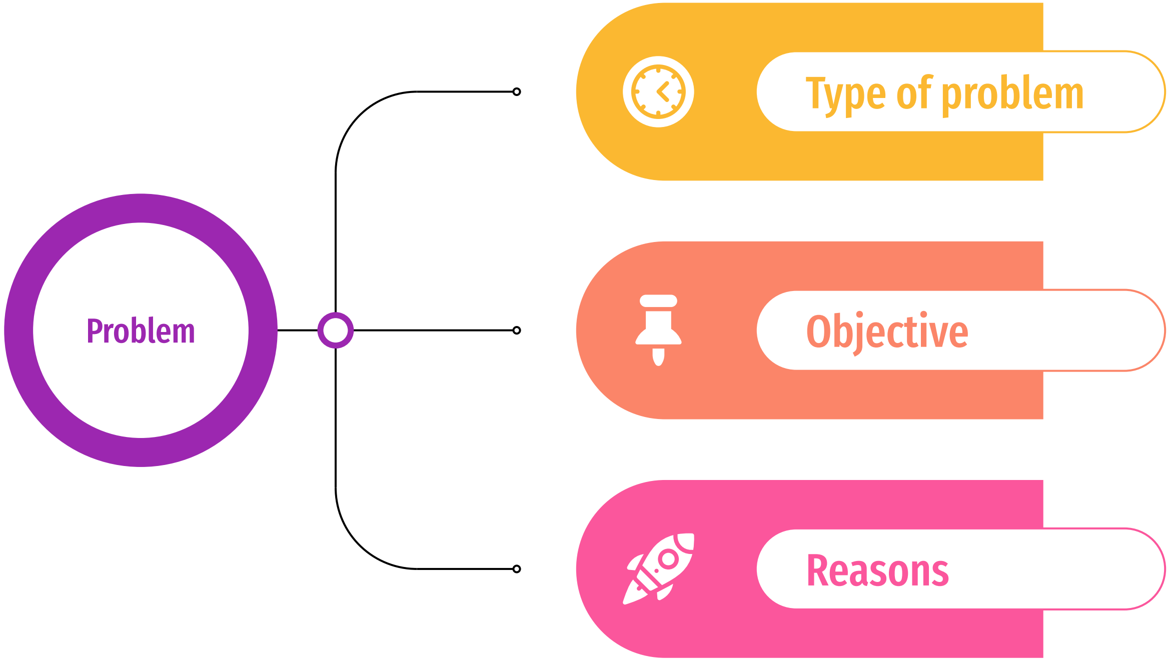
Feature Engineering

Encode categorical features

Modeling

Create model, train and test model

Define problem



Problem

Type of problem

Regression

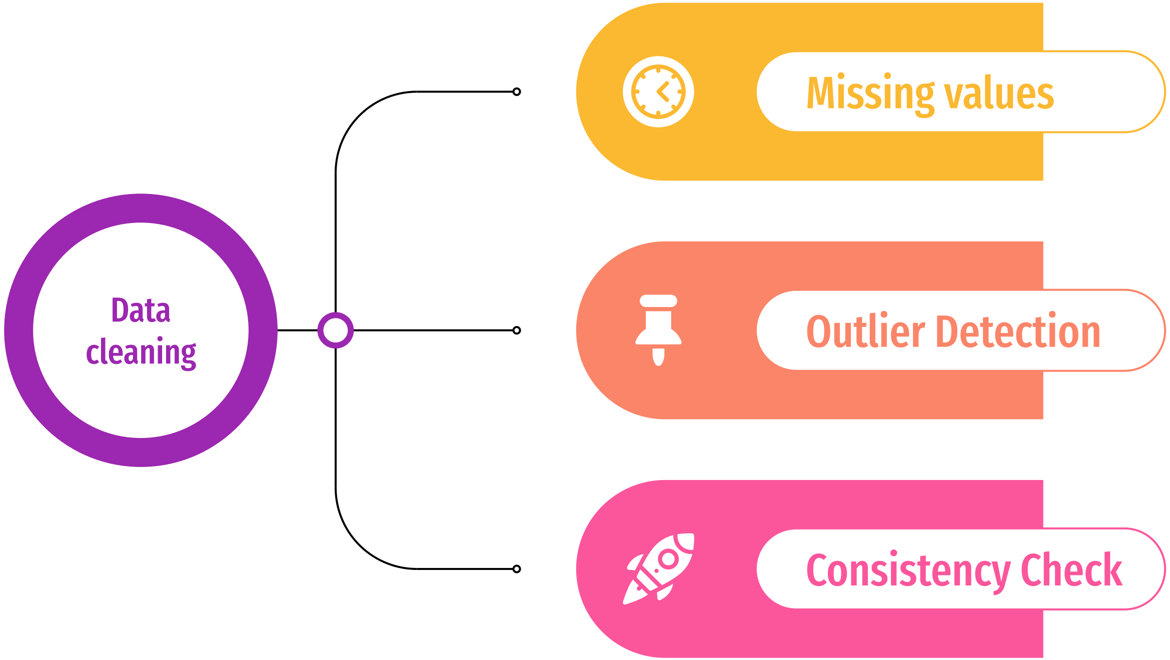
Objective

The total compensation of an employee based on various job-related features

Reasons

predict a continuous value, specifically the total compensation of an employee

Data Preparing

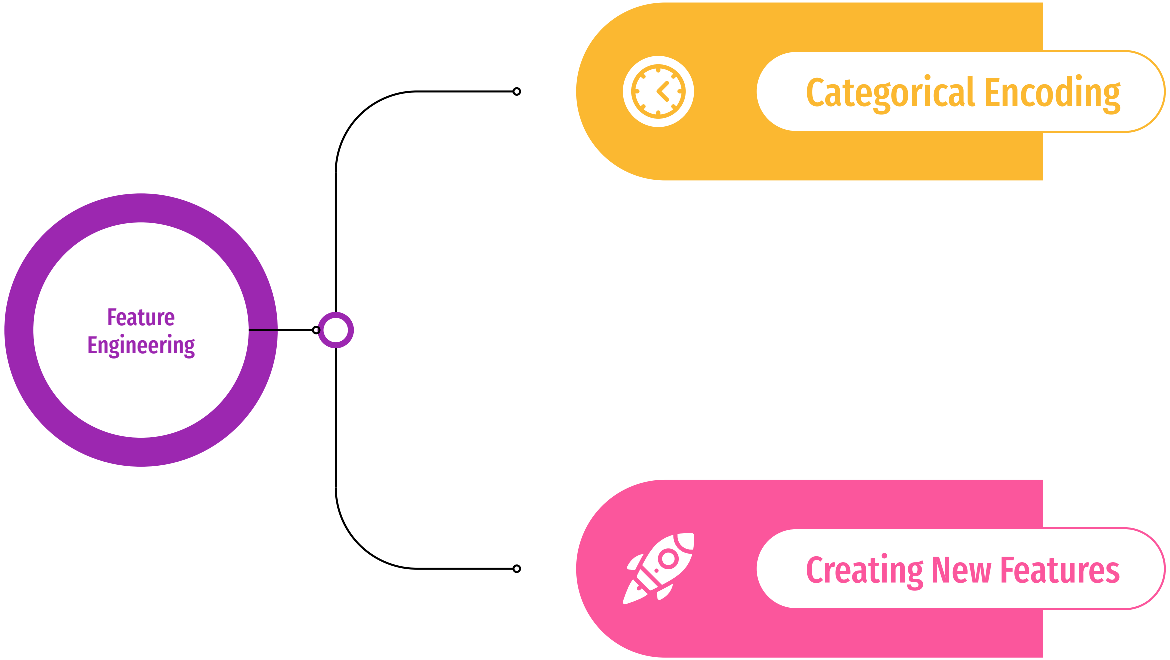


Replace them with the mean or mode

Calculate z-score and replace the outlier with the mean

Set all negative values to zero

Feature Engineering



Feature Engineering

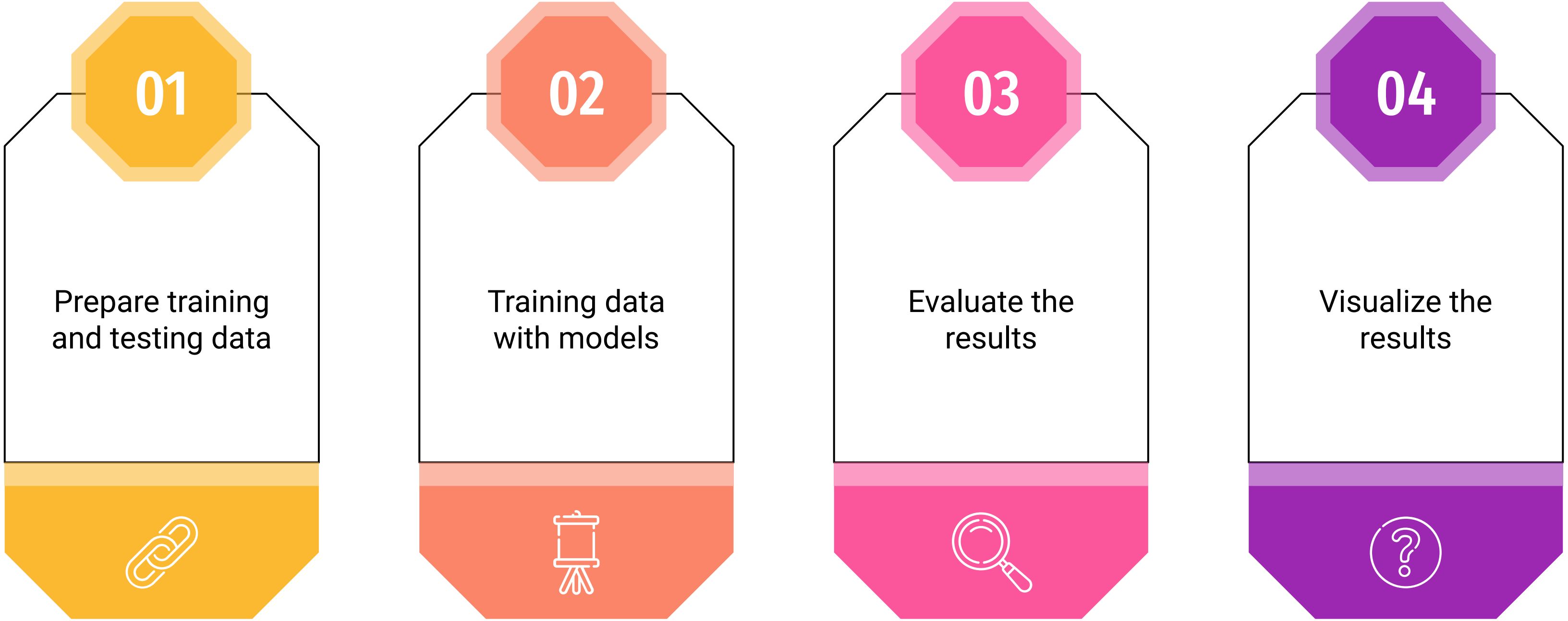
Categorical Encoding

Encode categorical columns

Creating New Features

Provide more information to train and improve the performance of the model.

Modeling



01

Prepare training and testing data

02

Training data with models

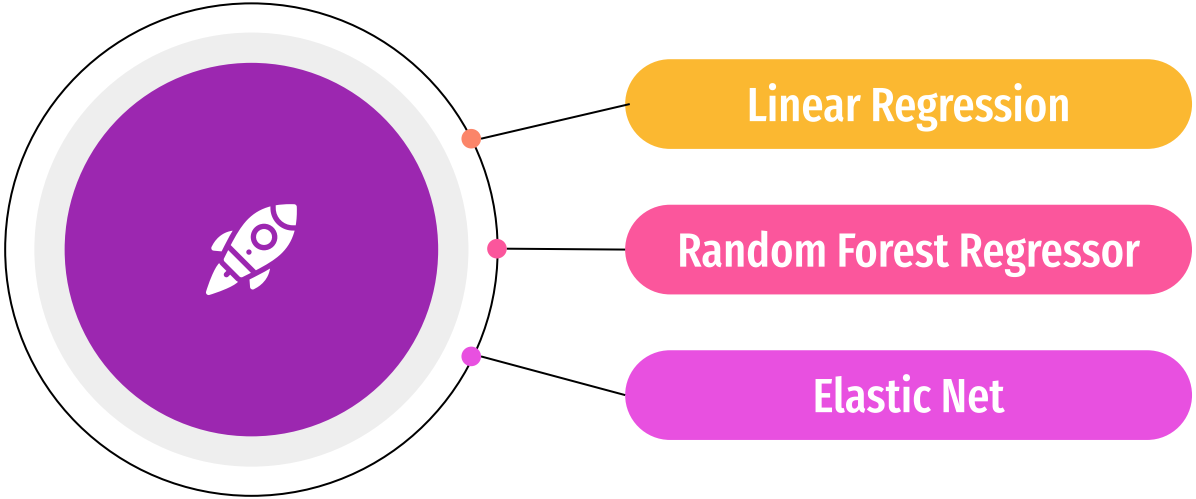
03

Evaluate the results

04

Visualize the results

Modeling



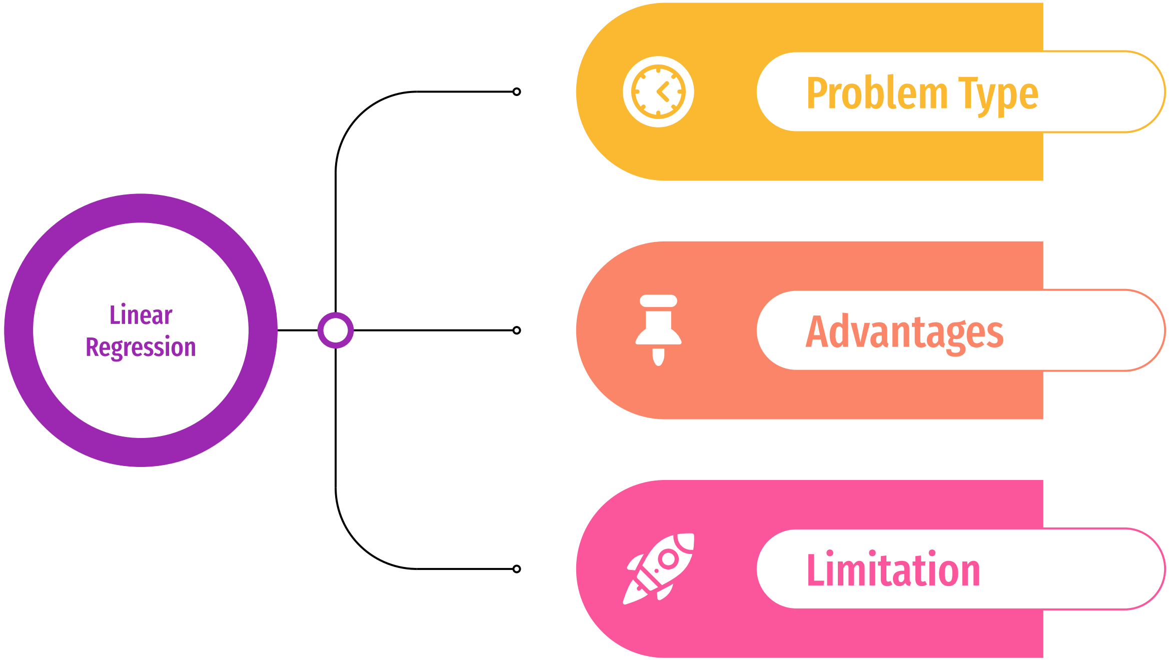
Linear Regression

Random Forest Regressor

Elastic Net

Models

Modeling

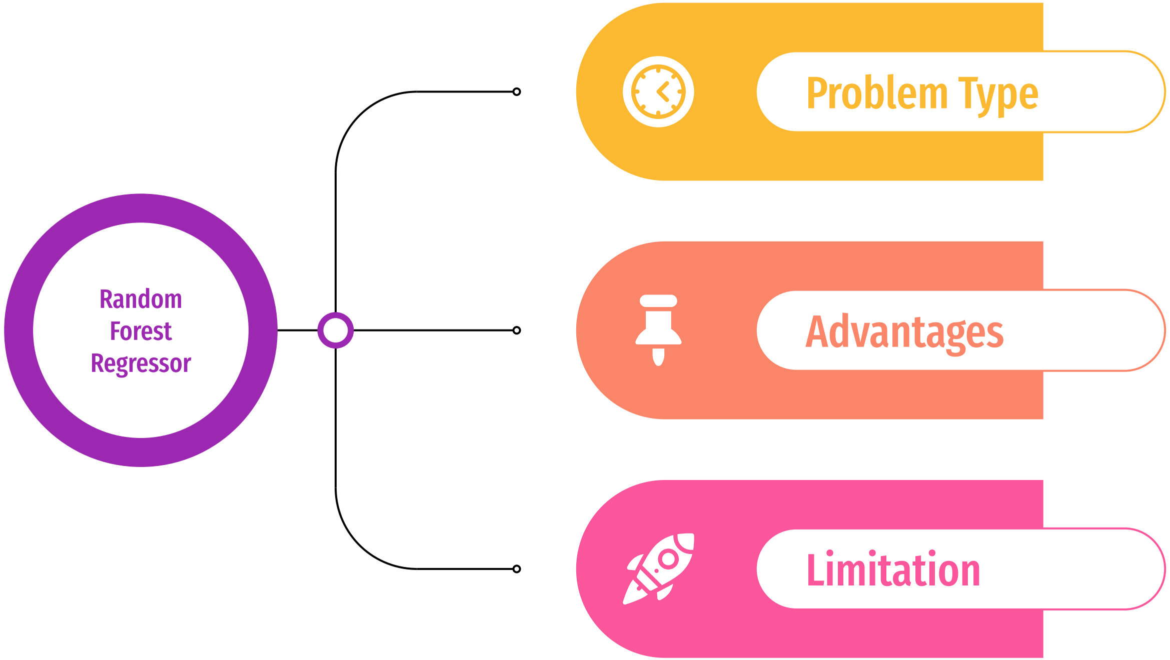


linear relationship between the (in)dependent variables

Simple, easy, and effective when the relationship is linear

handle non-linear relationships and complex

Modeling



Random Forest Regressor

Problem Type

both regression and classification

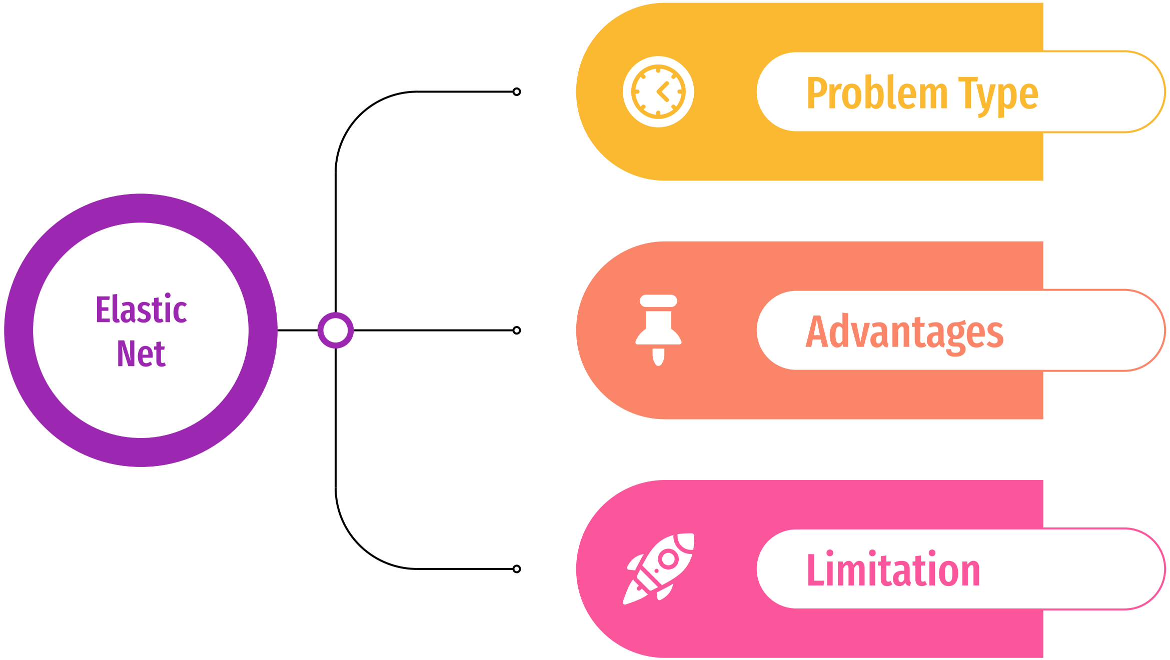
Advantages

Flexible, less sensitive to noise and overfitting.

Limitation

Requires a sufficiently large dataset to be effective.

Modeling



Elastic Net

Problem Type

linear relationship and high correlation between features

Advantages

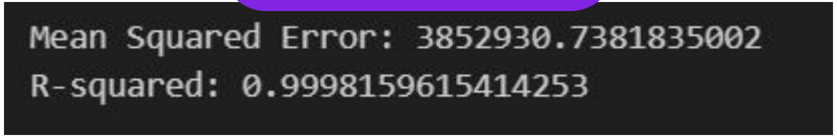
reducing overfitting and retaining important features

Limitation

May require tuning of hyperparameters for optimal performance

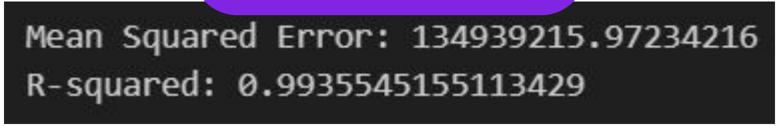
Compare the results

Linear Regression



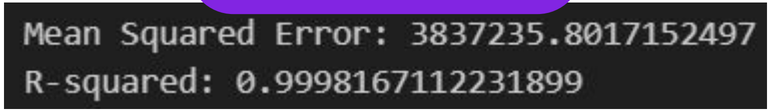
```
Mean Squared Error: 3852930.7381835002
R-squared: 0.9998159615414253
```

Random Forest



```
Mean Squared Error: 134939215.97234216
R-squared: 0.9935545155113429
```

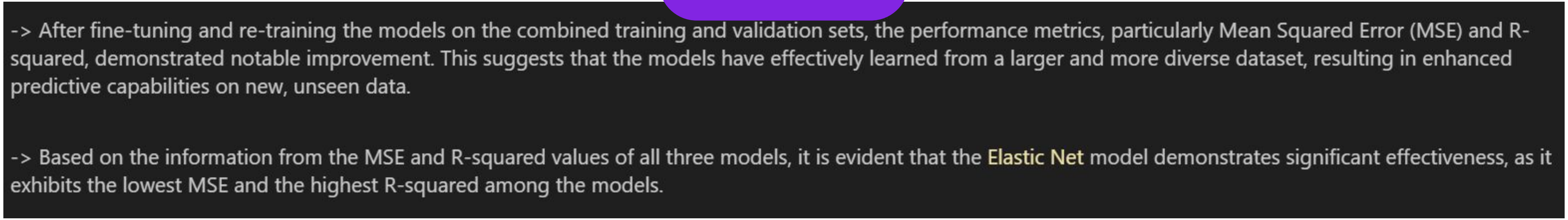
Elastic Net



```
Mean Squared Error: 3837235.8017152497
R-squared: 0.9998167112231899
```

Evaluate and compare

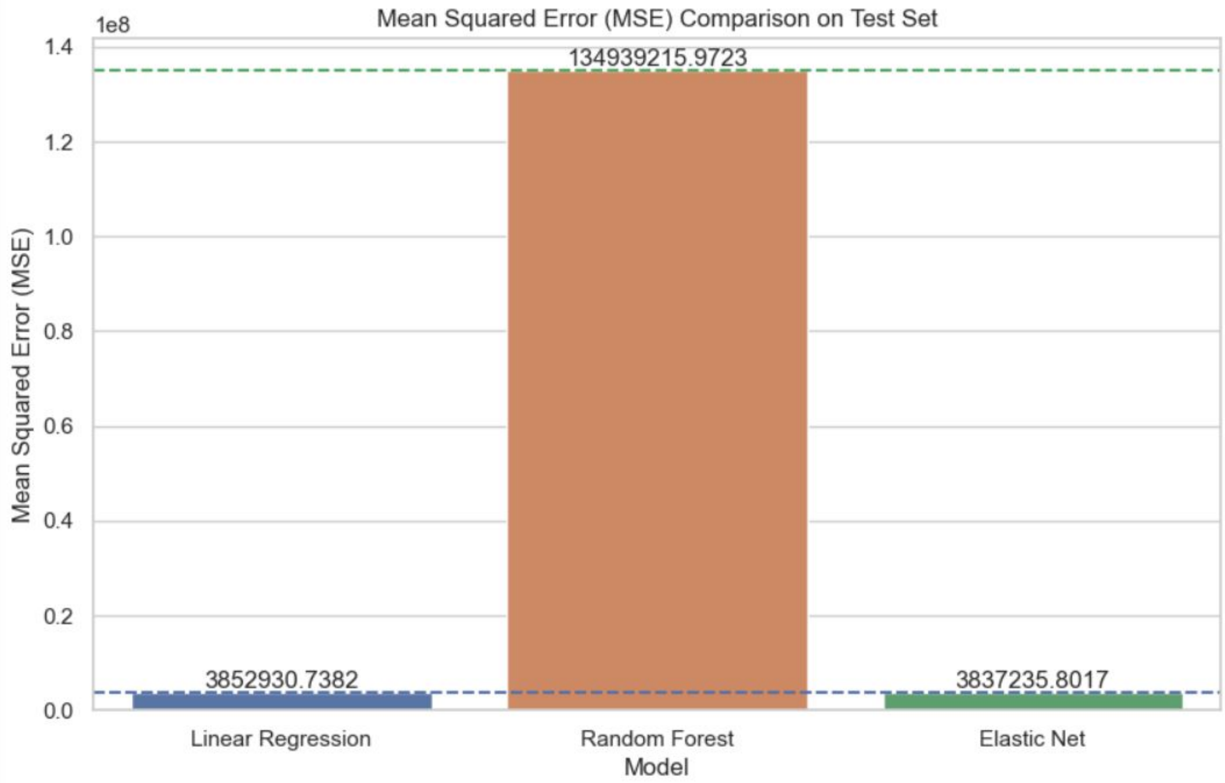
Conclusion



-> After fine-tuning and re-training the models on the combined training and validation sets, the performance metrics, particularly Mean Squared Error (MSE) and R-squared, demonstrated notable improvement. This suggests that the models have effectively learned from a larger and more diverse dataset, resulting in enhanced predictive capabilities on new, unseen data.

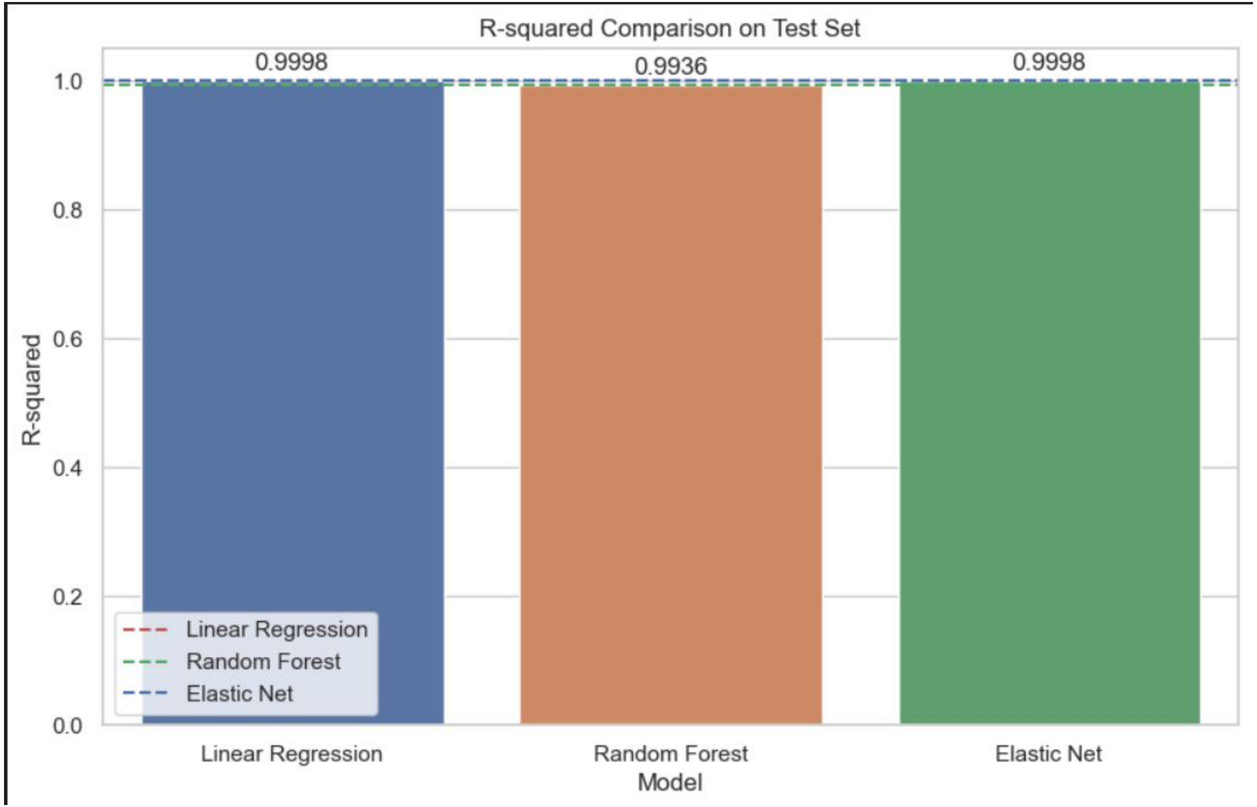
-> Based on the information from the MSE and R-squared values of all three models, it is evident that the Elastic Net model demonstrates significant effectiveness, as it exhibits the lowest MSE and the highest R-squared among the models.

Visualization



Mean Squared Error (MSE) Comparison on Test Set

Visualization



Thanks for your listening!