



STEM Salary Prediction

資科碩一	110753140	張立暘
資科三	108703017	邱彥翔
資科碩一	110753163	林昱辰
社會二	109204035	黃楷捷

Table of contents

01 Motivation

You can describe
the topic of the
section here

02 Dataset intro.

You can describe
the topic of the
section here

03 Model Experiment

You can describe
the topic of the
section here

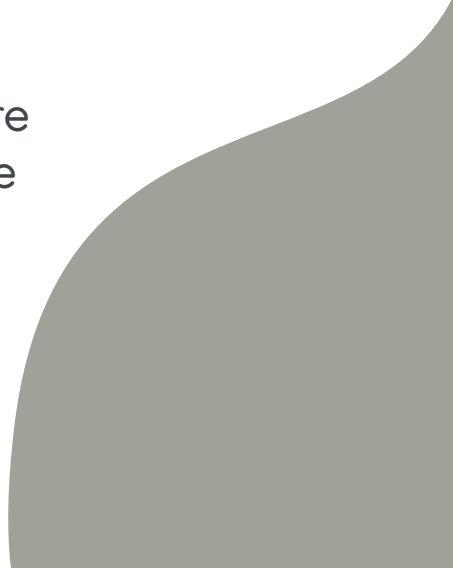
04 Comments

You can describe
the topic of the
section here

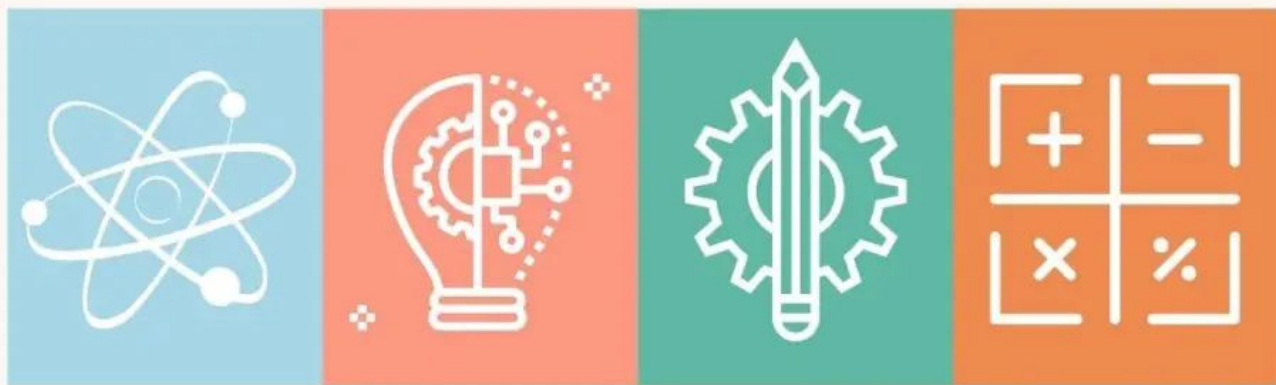


Motivation

To help limit distractions or make video calls more fun, you can now blur your background or replace your background with an image



STEM



Science·Technology·Engineering·Mathematics





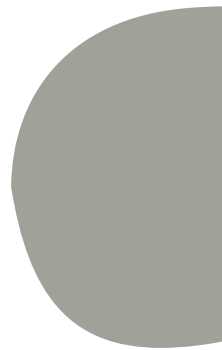
?



?



?



The background features several abstract elements: a large, light beige organic blob in the center; a smaller, brownish-orange organic blob in the upper right; a grey circle on the left partially covered by a hexagonal grid pattern; and a thin black horizontal line below the title. In the bottom right corner, there is a simple line drawing of a plant branch with several leaves.

Dataset Introduction

Dataset Resource

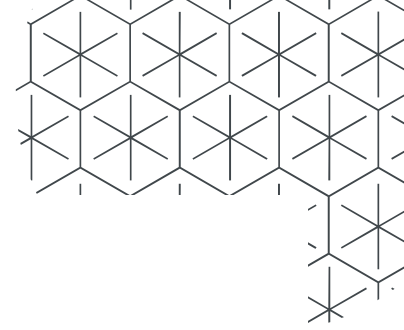


<https://www.kaggle.com/jackogozaly/data-science-and-stem-salaries>

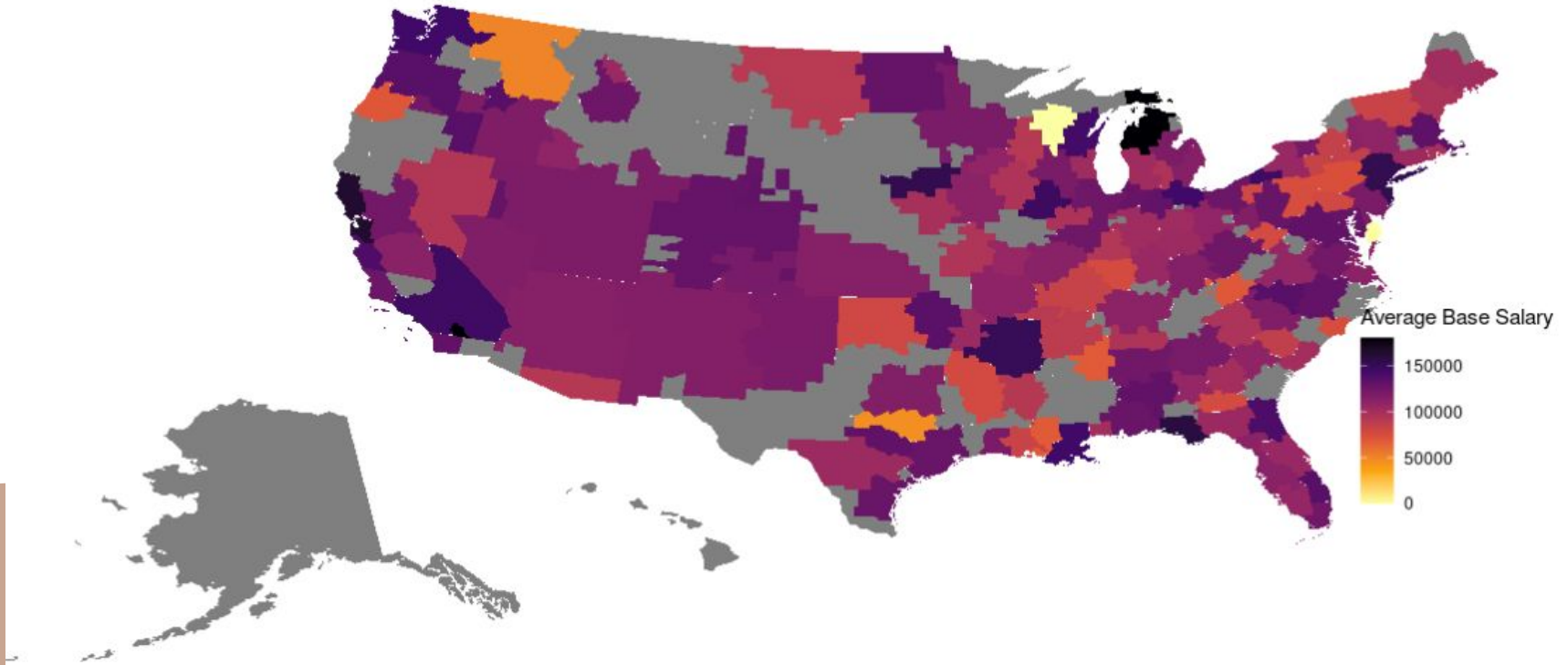
Feature	Type	Description	Example
timestamp	Nominal	時間戳記	6/7/2017 11:33:27
company	Nominal	所任職的公司	Oracle
level	Nominal	職業層級	L3
title	Nominal	職稱	Product Manager
totalyearlycompensation	Numeric	總體薪酬	127000
location	Nominal	公司所在地區	Redwood City, CA
yearsofexperience	Numeric	從事這行業的時長	1.5
yearsatcompany	Numeric	在這公司任職的時長	1.5
tag	Nominal	職業屬性別	API Development (Back-End)

Feature	Type	Description	Example
basesalary	Numeric	基本薪水	107000
stockgrantvalue	Numeric	股票分紅	20000
bonus	Numeric	獎金	10000
gender	Nominal	性別	Male
otherdetails	Nominal	備註	New grad offer -- intern conversion
cityid	Numeric	公司所在地區編號	7392
dmaid	Numeric	地區區碼	807
rowNumber	Numeric	編號	1
Race	OneHotEncoding	種族	White
Education	OneHotEncoding	教育程度	Master's Degree

Exploratory Data Analysis

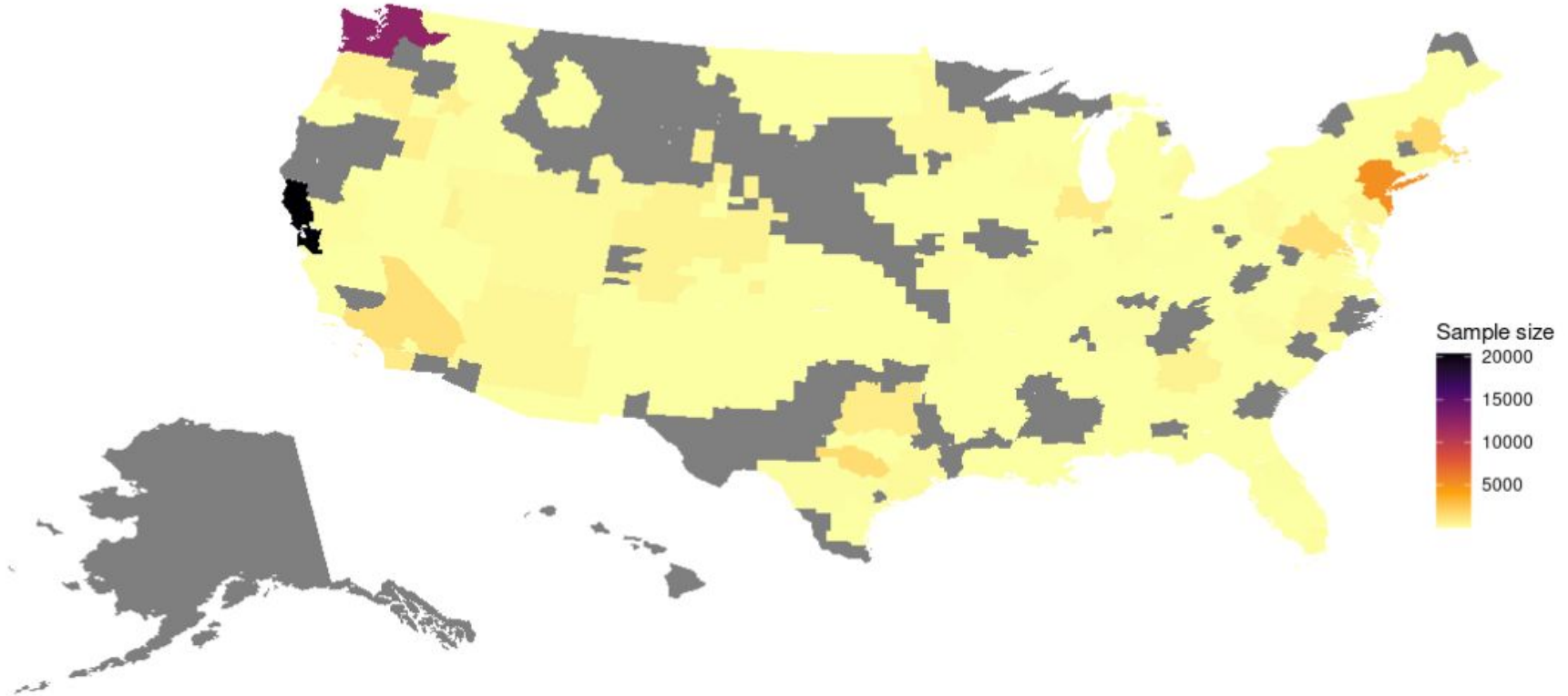


Average base salary by Designated Market Area (DMA)

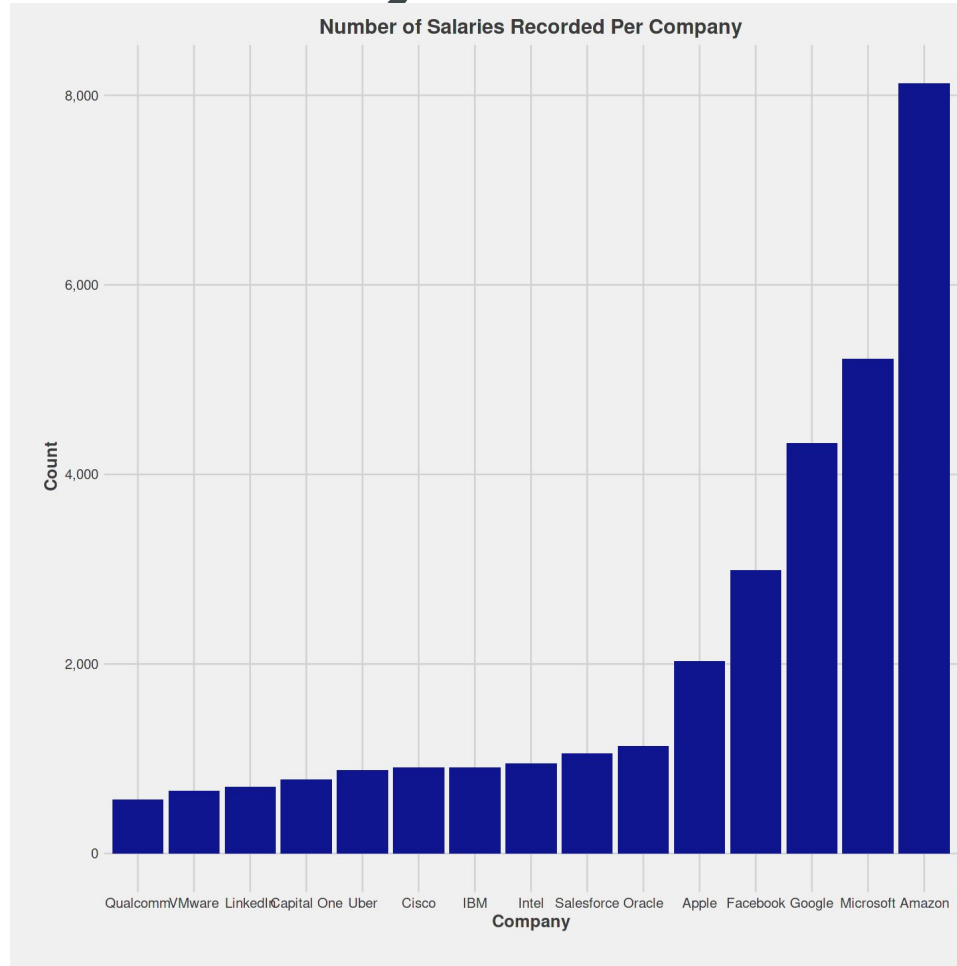


Exploratory Data Analysis

Number of sample by Designated Market Area (DMA)

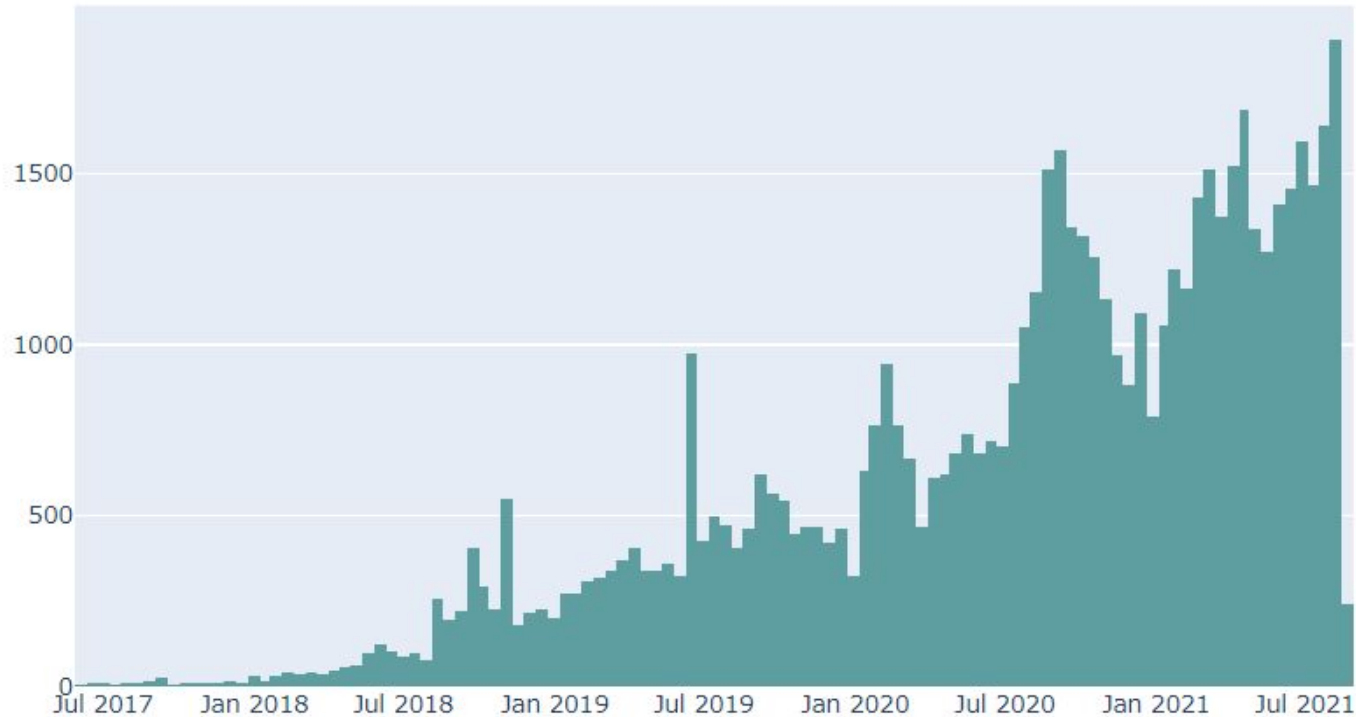


Exploratory Data Analysis

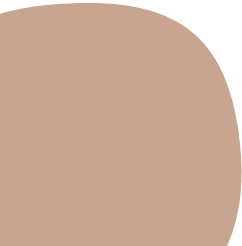
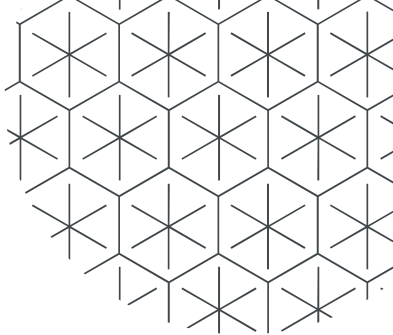
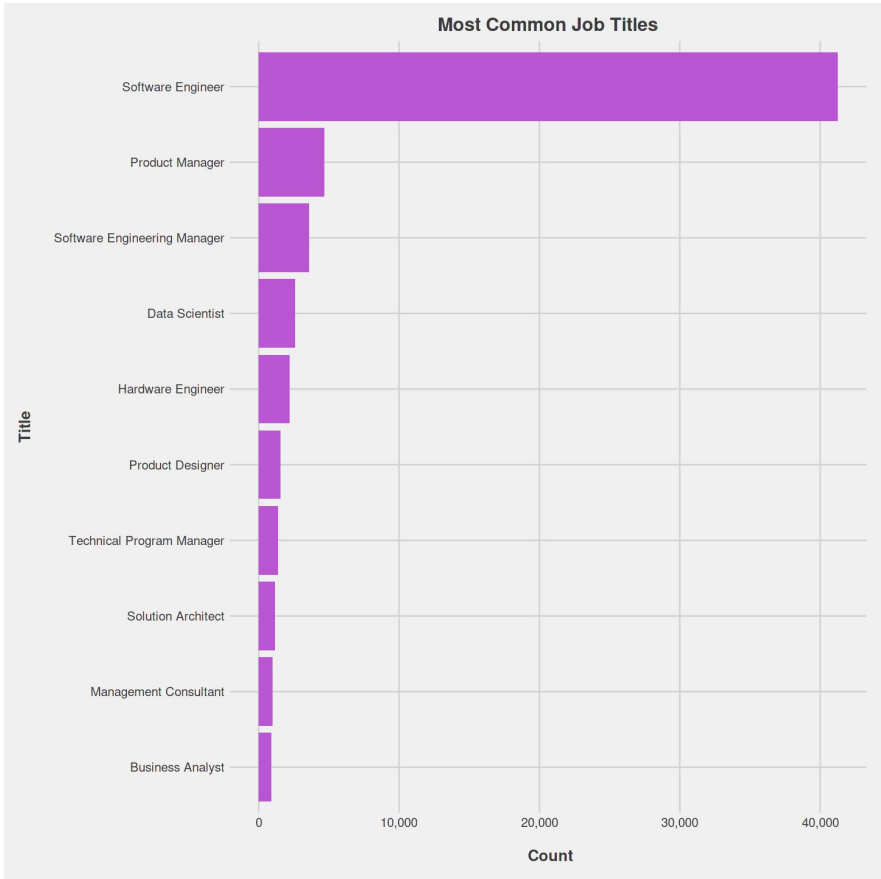


Exploratory Data Analysis

STEM jobs in time

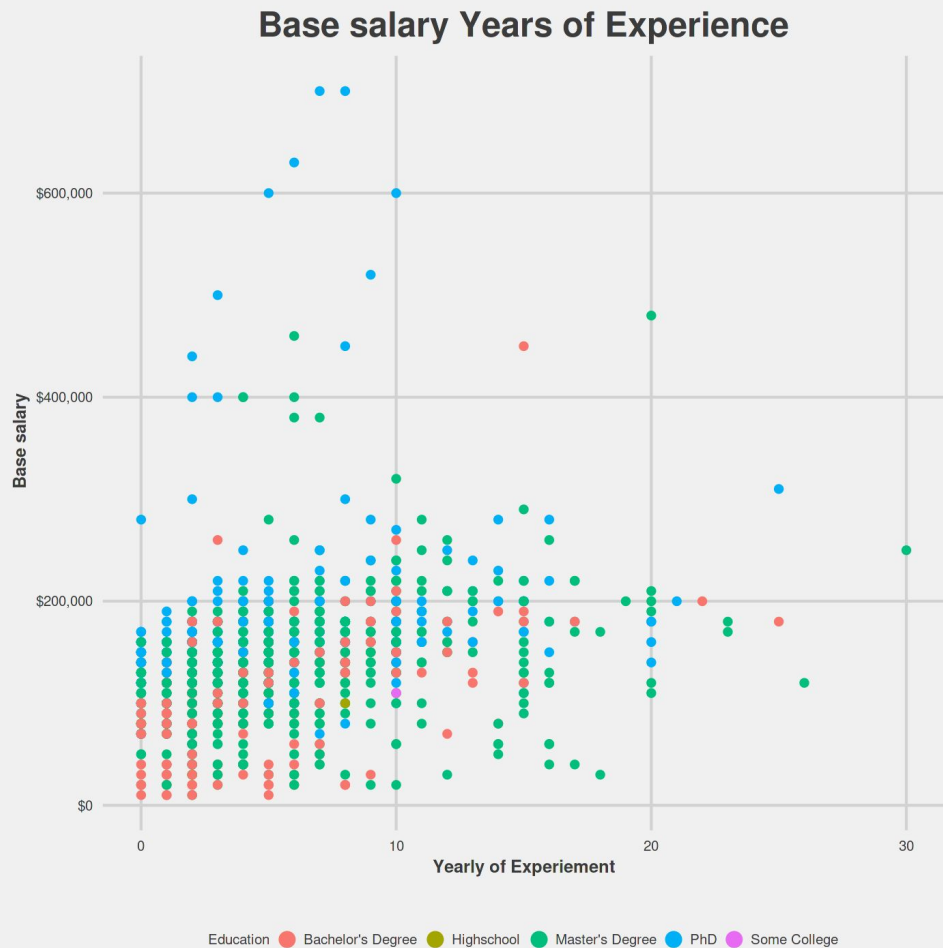


Exploratory Data Analysis



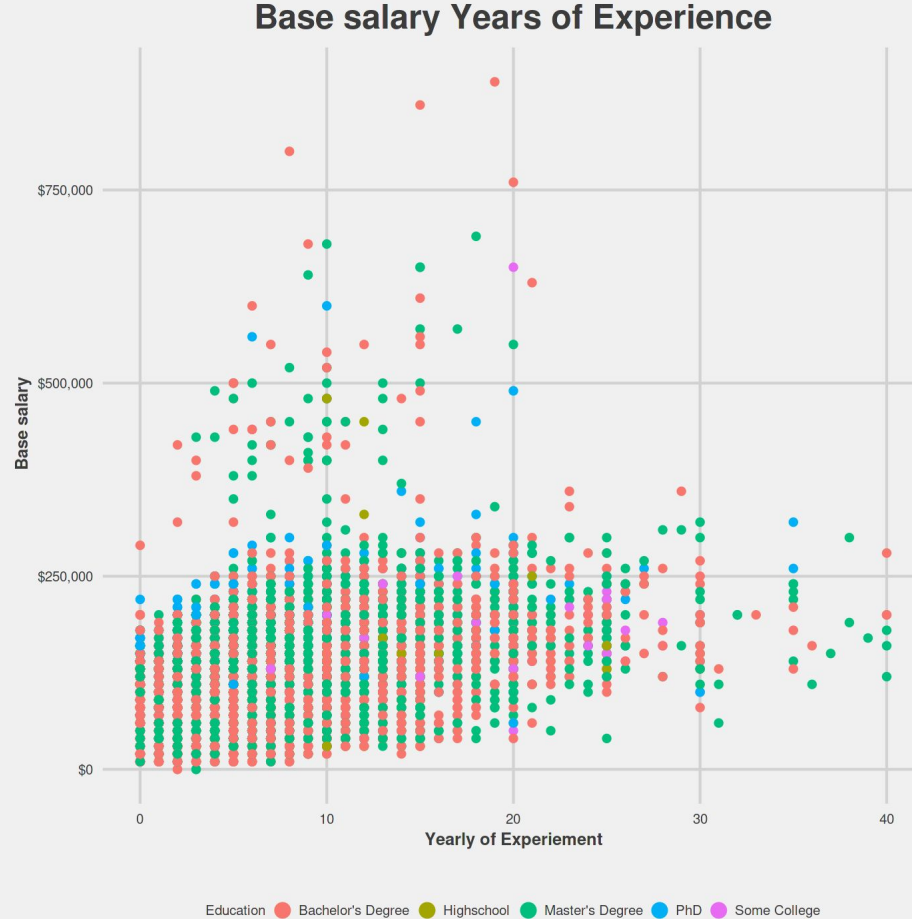
Exploratory Data Analysis

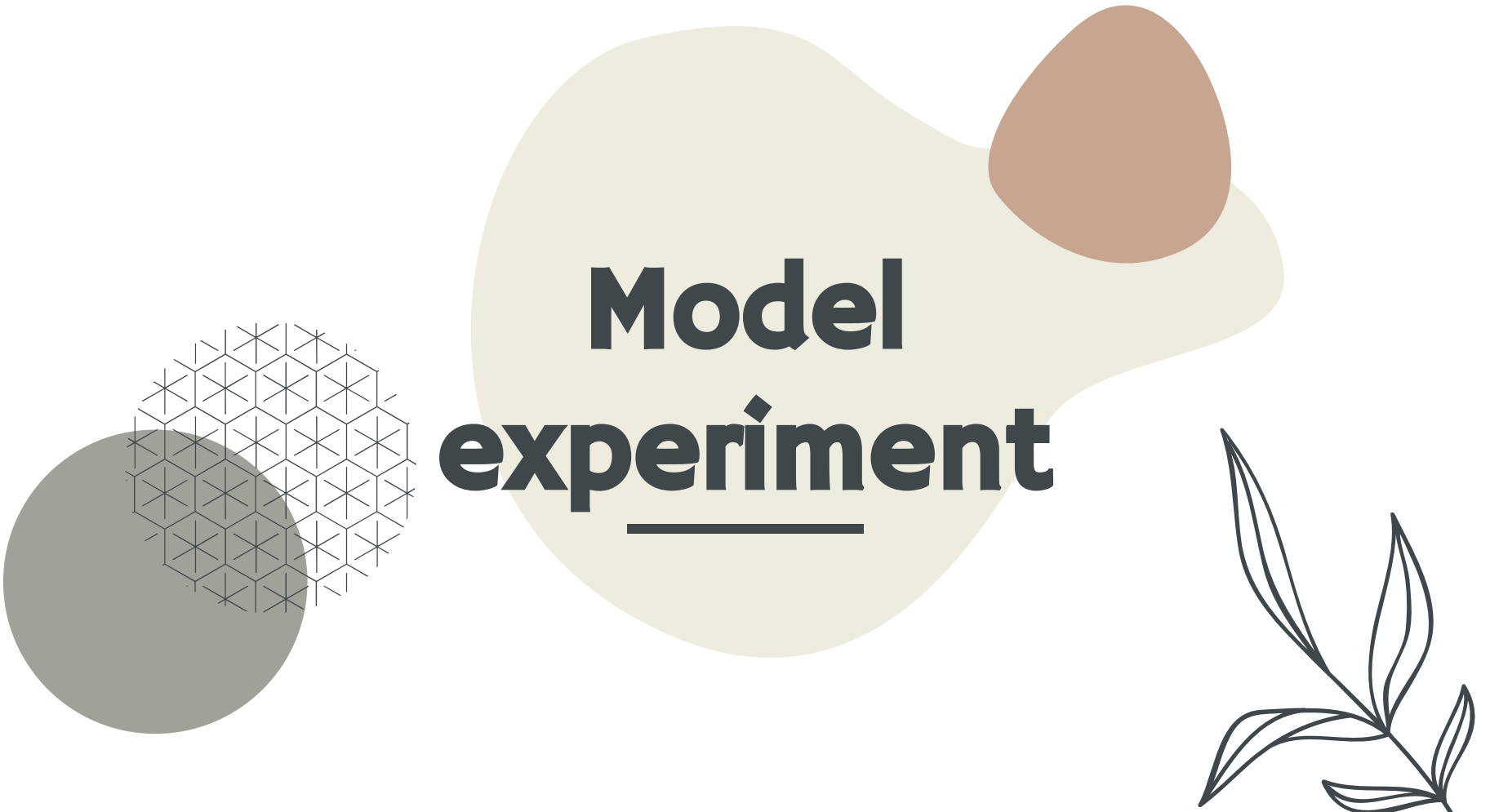
Data Scientist



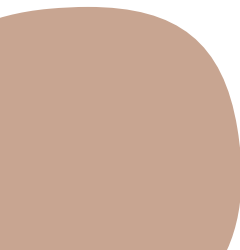
Exploratory Data Analysis

Software Engineer

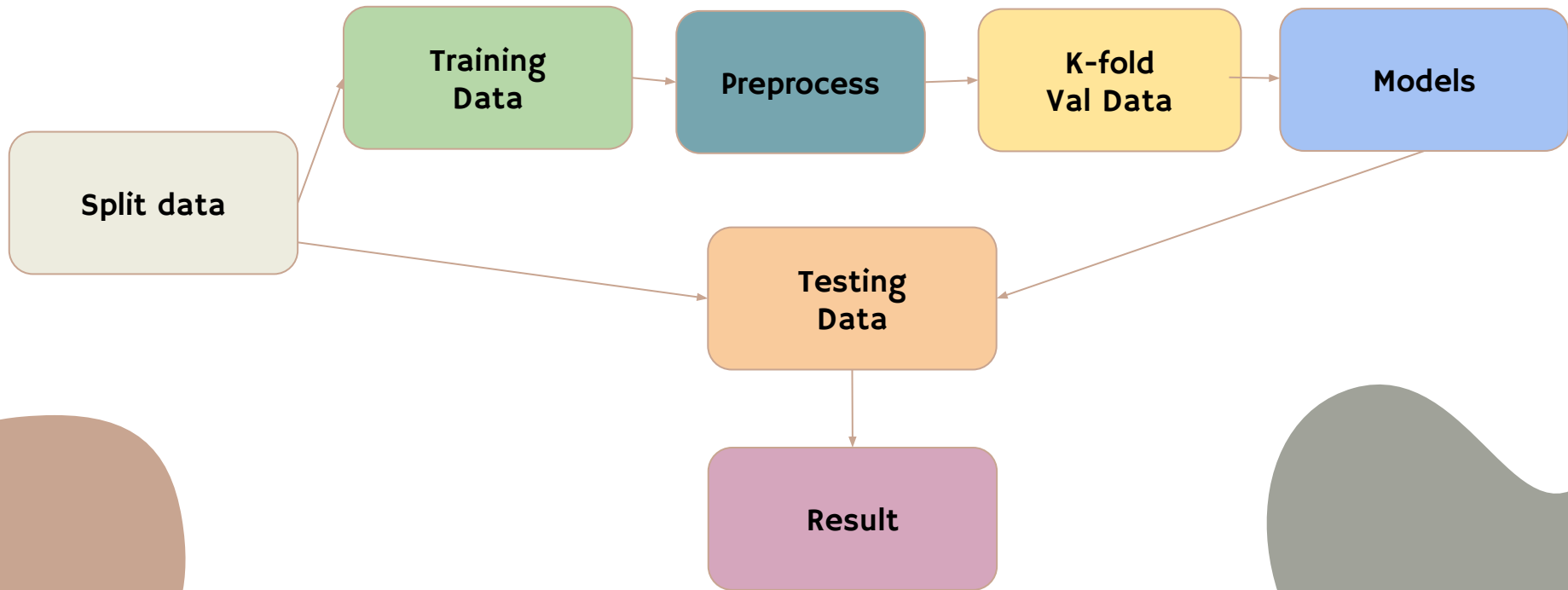


The background features several abstract elements: a large, light beige organic blob in the center; a smaller, brownish-orange organic blob in the upper right; a grey circle on the left partially covered by a hexagonal grid pattern; and a simple line drawing of a plant branch with three leaves in the bottom right corner.

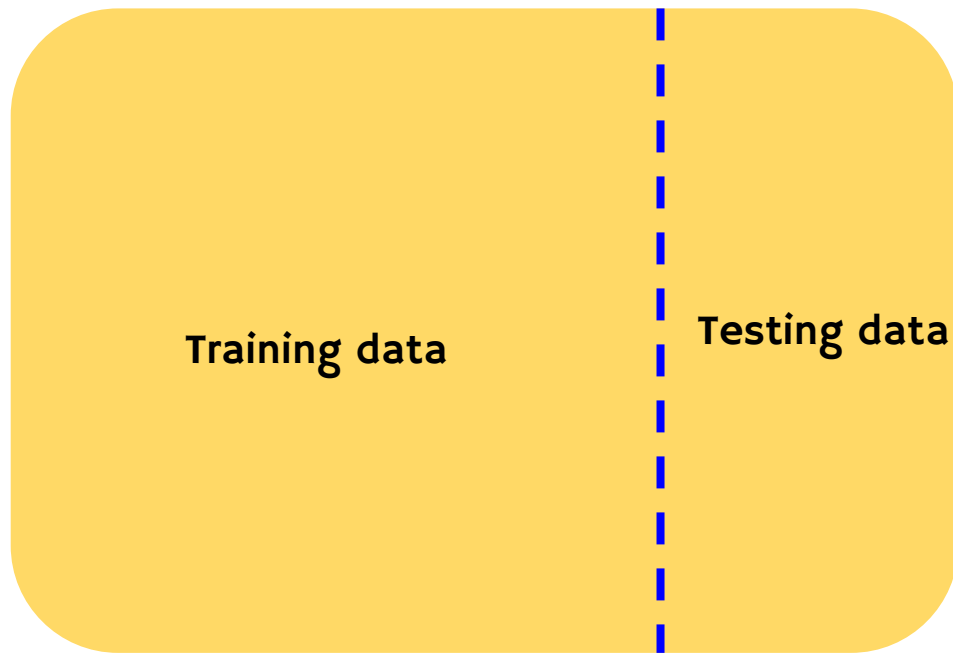
Model experiment



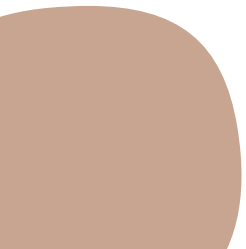
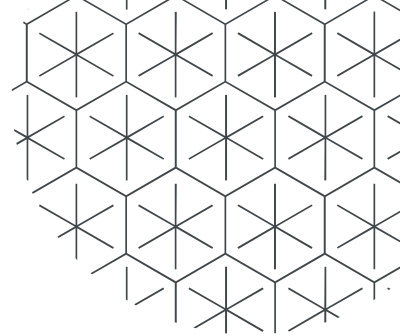
Data Pipeline



Split dataset

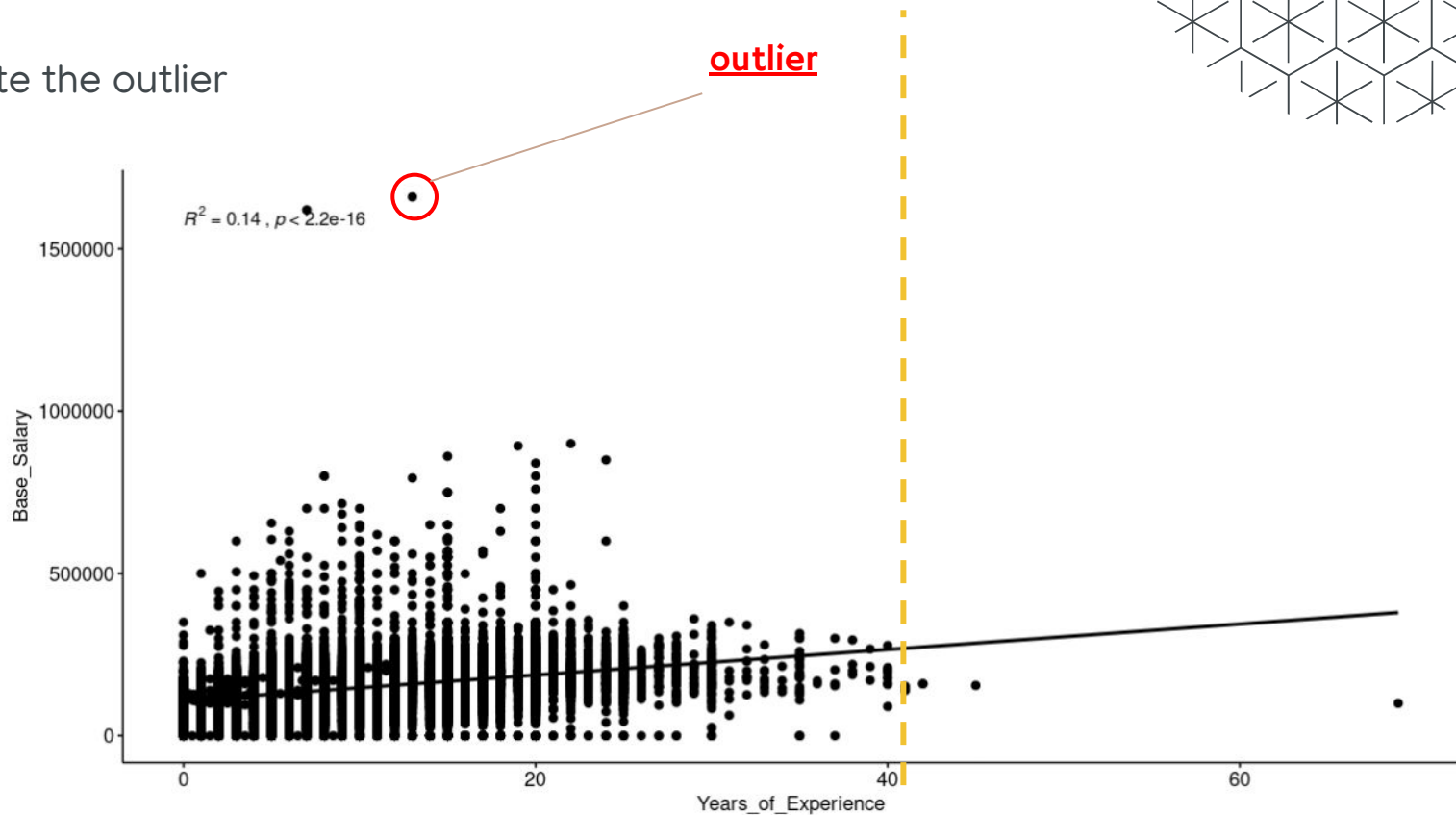


Original data



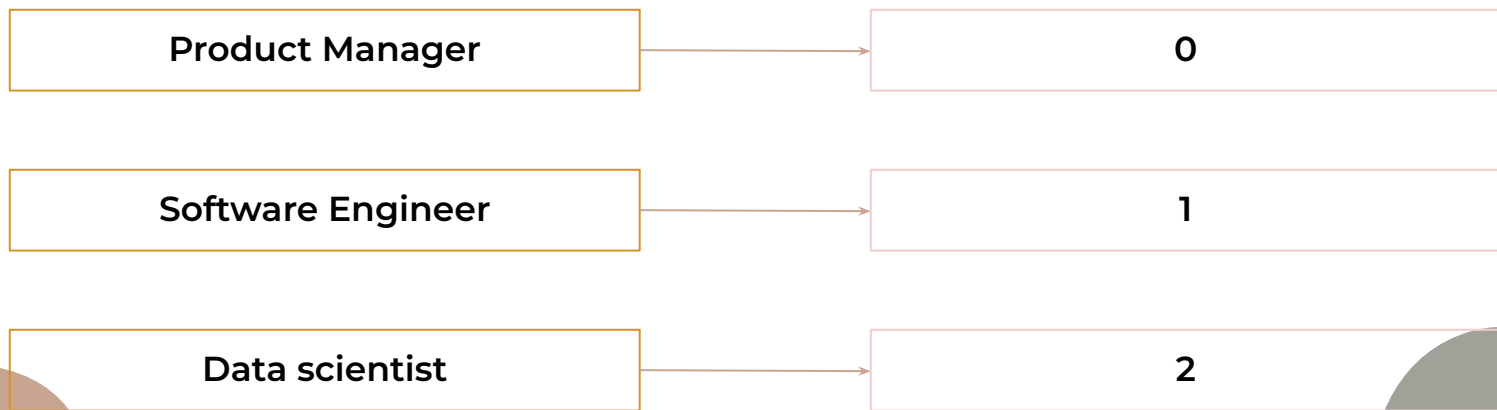
Preprocess

- Delete the outlier



Preprocess

- Label encoding categorical data



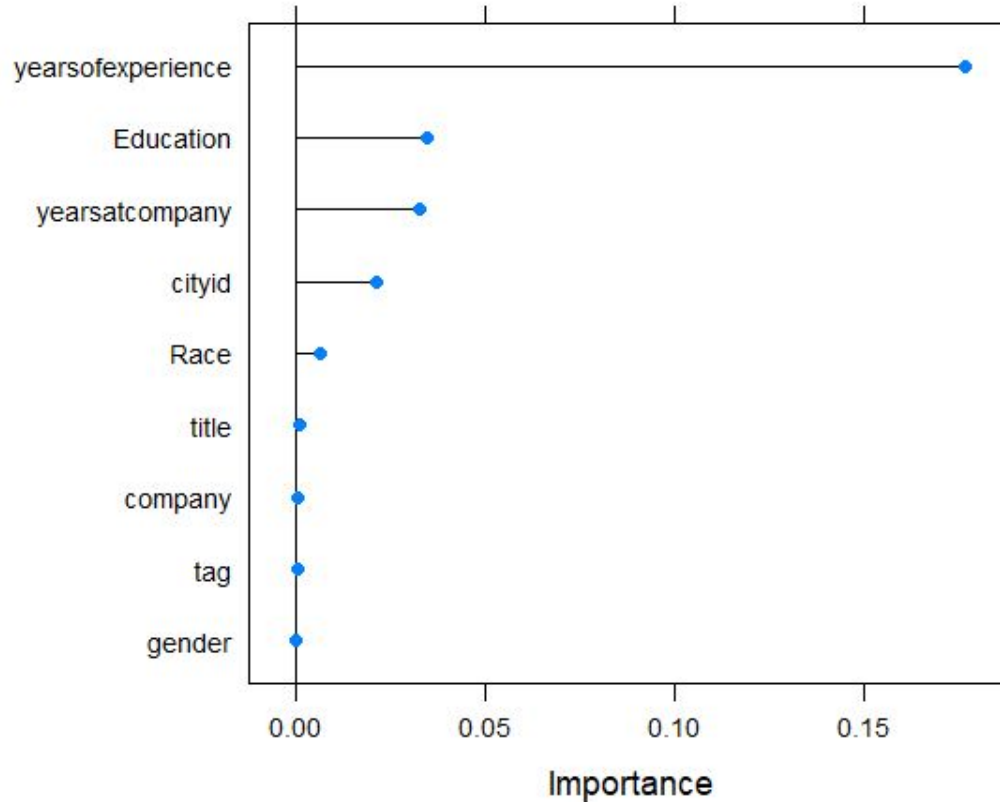
Performance

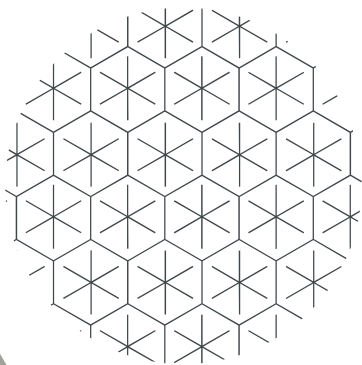
	Valid MAE	Test MAE
NULL MODEL	41651.24	41809.7
Linear Regression	3619013	35596.49
SVM	35923.4.9	35614.69
Decision Tree	33123.33	34400.99
Random forest	19671.20	21822.02
XGBtree	18185.23	20744.34

10-Fold Performance

	Valid MAE	Test MAE
Linear Regression	41651.24	35596.49
SVM	35921.92	35614.69
Decision Tree	33123.33	34400.99
Random forest	19473.32	21822.02
XGBtree	18048.19	20677.39

Variable Importance



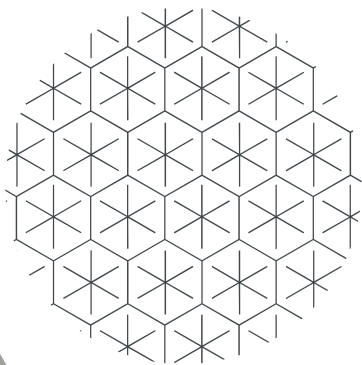


DEMO





<https://yhqchiu.shinyapps.io/code/>

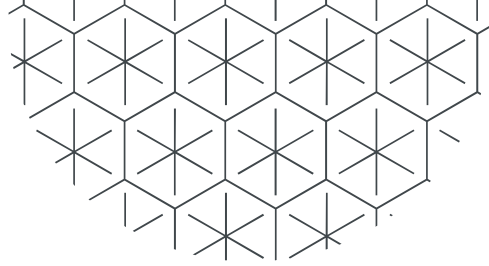


Comment



Challenging points

- NA值很多，嘗試用KNN來補效果卻不大好
- Shiny app 呈現會有一些大小的問題，以及無法正確visualization
- data science的project分工以及merge code是一個大問題



**Thank for
listening !**