CREDIT SCORE

Data Analysis

Ying Cheung
Elias Grant
Ivan Chuang
Lea Choe
Manshu Huang
Maya Thompson

NSDC SPRING 2025

Project Overview

Objectives:

1. Create visualizations to explore relationships between variables in **`users_data.csv`** from the Financial Transactions dataset.

2. Analyze **`users_data.csv`** and identify a valid model to predict credit score from various explanatory variables.

DATA ANALYSIS

Credit Score

Data Collection

Week 3 (data source, number of files, variables, variable types)

- Focused on two files: cards_data.csv and users_data.csv
- Variables (card_data)
  - Card_id
  - User_id
  - Card_limit
  - Card_type
  - issued_date
- Variables (users_data)
  - User_id
  - Current_age
  - Retirement_age
  - Gender
  - Per_capita_income
  - Yearly_income
  - Total_debt
  - Num_credit_cards
  - Birth_year / birth_month

Data source: Kaggle Dataset: Transactions Fraud Datasets

id ID	client_id	card_brand	card_type	card_number	expires	cvv
0 – 6145	0 – 1999	Mastercard 52% Visa 38% Other (611) 10%	Debit 57% Credit 33% Other (578) 9%	300106b – 6997197b	1997-06-30 – 2024-11-30	0
4524	825	Visa	Debit	4344676511950444	12/2022	623
2731	825	Visa	Debit	4956965974959986	12/2020	393
3701	825	Visa	Debit	4582313478255491	02/2024	719
42	825	Visa	Credit	4879494103069057	08/2024	693
4659	825	Mastercard	Debit (Prepaid)	5722874738736011	03/2009	75
4537	1746	Visa	Credit	4404898874682993	09/2003	736
1278	1746	Visa	Debit	4001482973848631	07/2022	972
3687	1746	Mastercard	Debit	5627220683410948	06/2022	48
3465	1746	Mastercard	Debit (Prepaid)	5711382187309326	11/2020	722
3754	1746	Mastercard	Debit (Prepaid)	5766121508358701	02/2023	908
5144	1718	Mastercard	Debit	5495199163052054	03/2022	677
2029	1718	Mastercard	Debit	5804499644308599	07/2023	258
2379	1718	Mastercard	Debit	5766352389579834	02/2020	992
2732	1718	Visa	Debit	4242015583697294	06/2020	928
4706	1718	Mastercard	Debit	5191030913182493	06/2024	360
281	708	Visa	Credit	4017261190134817	05/2015	877
1106	708	Mastercard	Debit (Prepaid)	5581970288727991	06/2020	448

```{r}
# MUTATE - permanently create card_brand_num -- quantitative version of
card_brand
# card_brand still exists
cards_data <-
  cards_data |>
  mutate(card_brand_num = recode(card_brand,
                                 "Amex" = 1,
                                 "Discover" = 2,
                                 "Mastercard" = 3,
                                 "Visa" = 4))
```

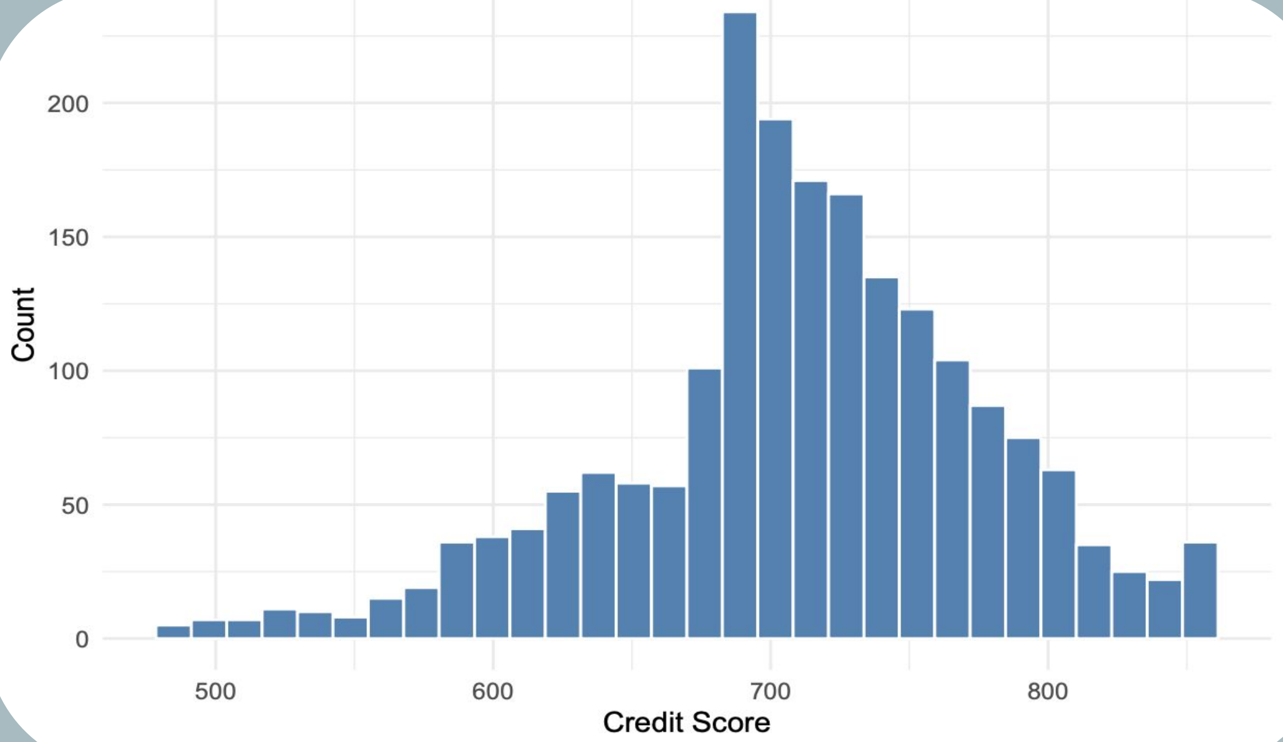
Data Cleaning

Week 3

- **Cards_data.csv:** Checked structure and content with str(), head ( ), tail ( )
  - Found no missing values and no duplicate rows
  - Transformed categorical variables into numeric formats (ex. Card_vrand → card_brand_num)
- **Users_data.csv:** used str( ), head ( ), tail ( )
  - Found no missing values and no duplicate rows
  - Converted categorical variables into numeric
    - Ex. Gender → gender_num

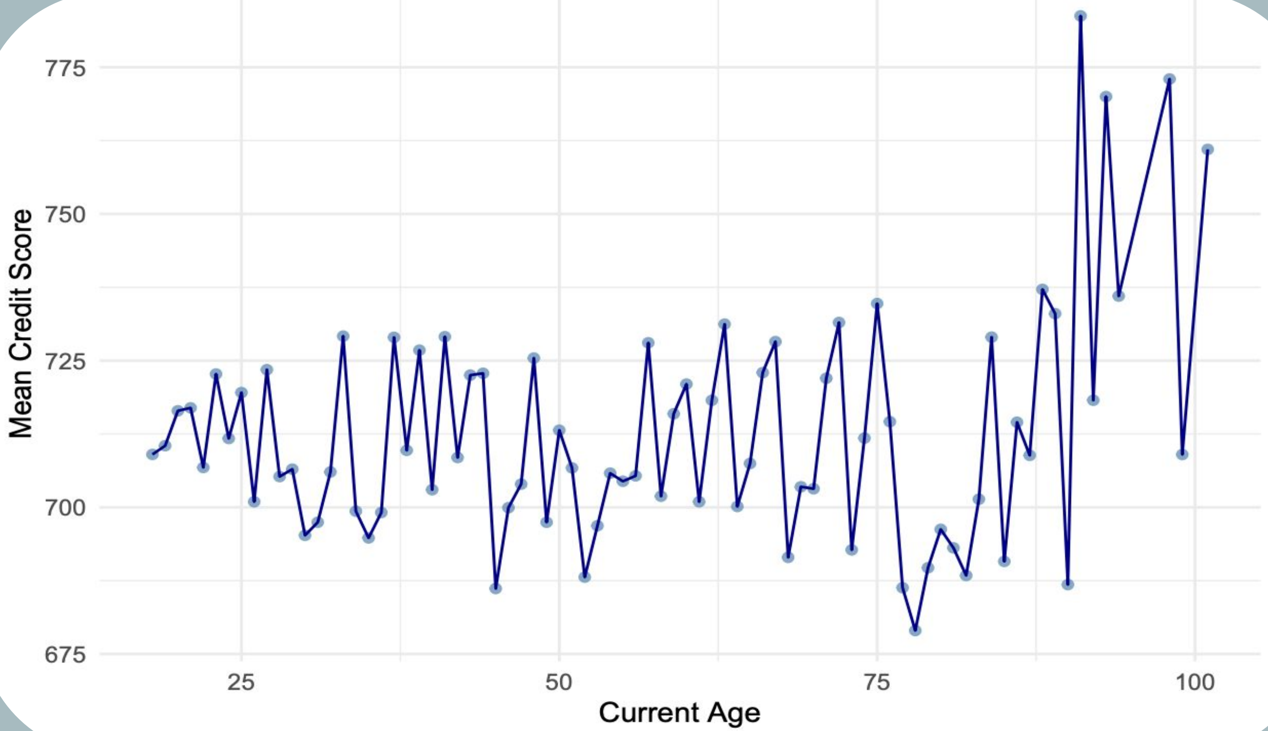
Exploratory Data Analysis

Distribution of Credit Scores



Exploratory
Data
Analysis

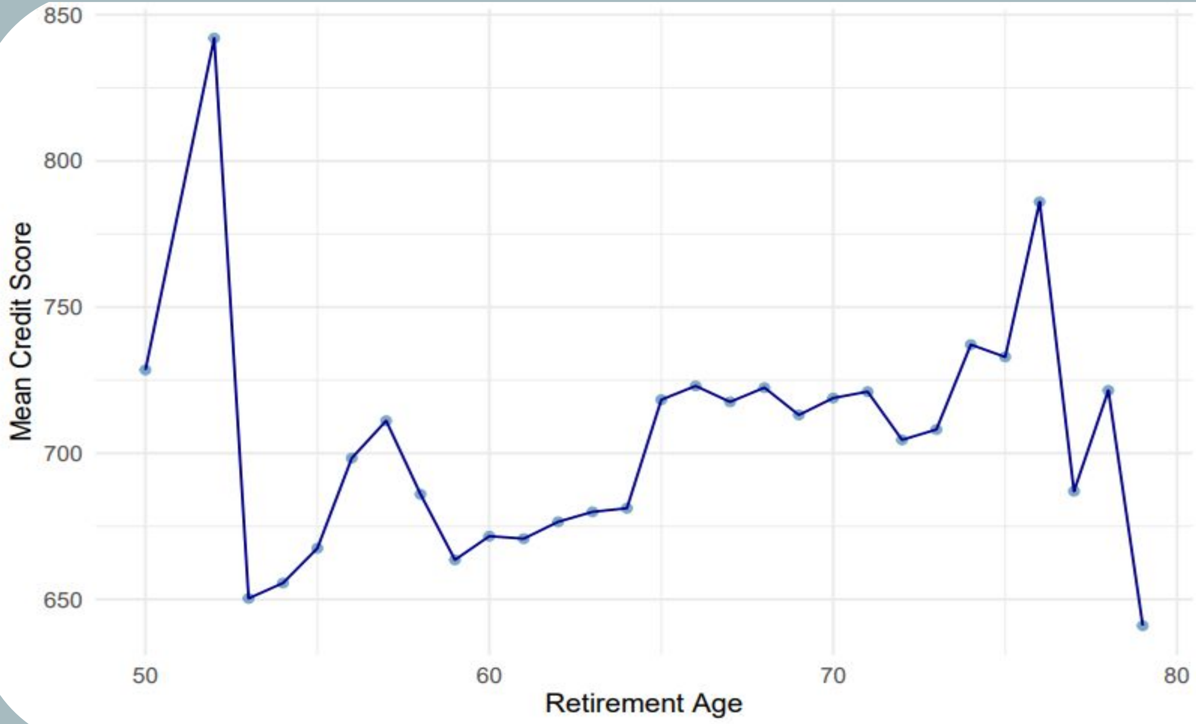
Average Credit Score by Current Age



Credit Score

Exploratory
Data
Analysis

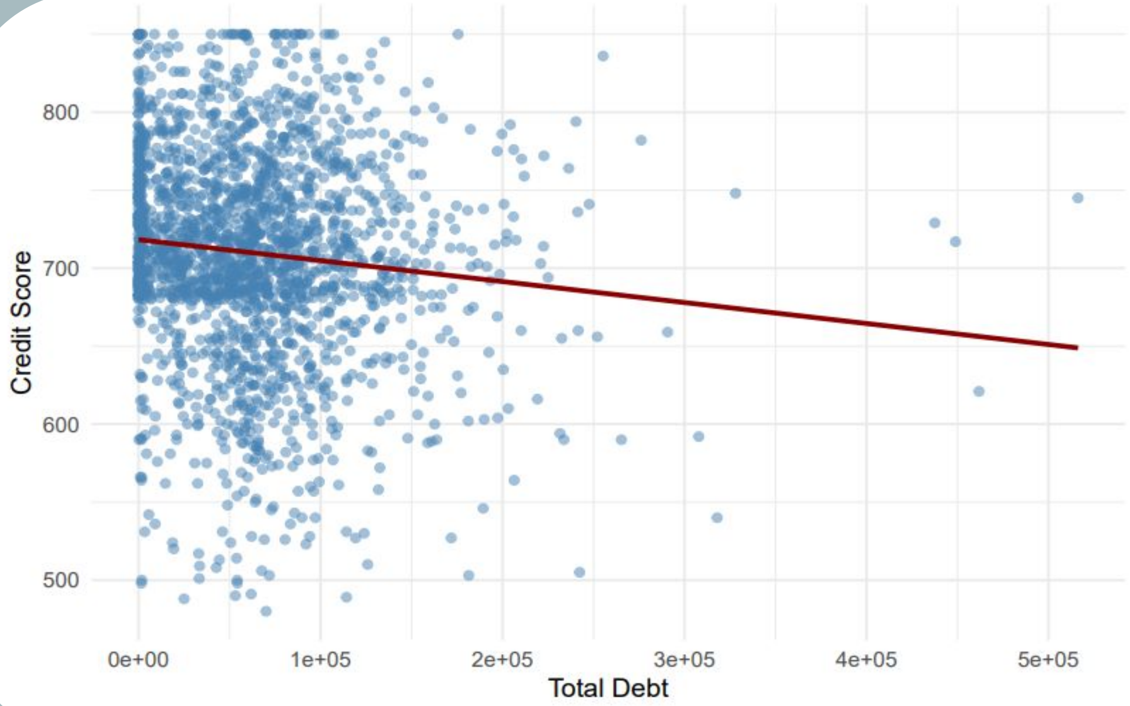
Average Credit Score by Retirement Age



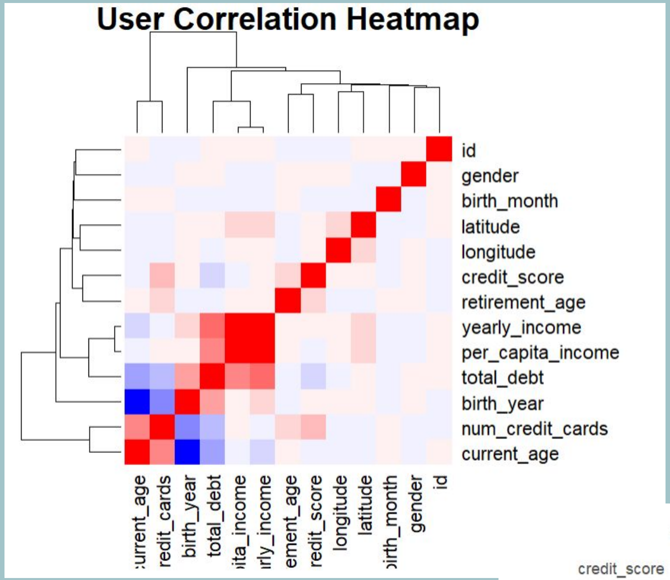
Credit Score

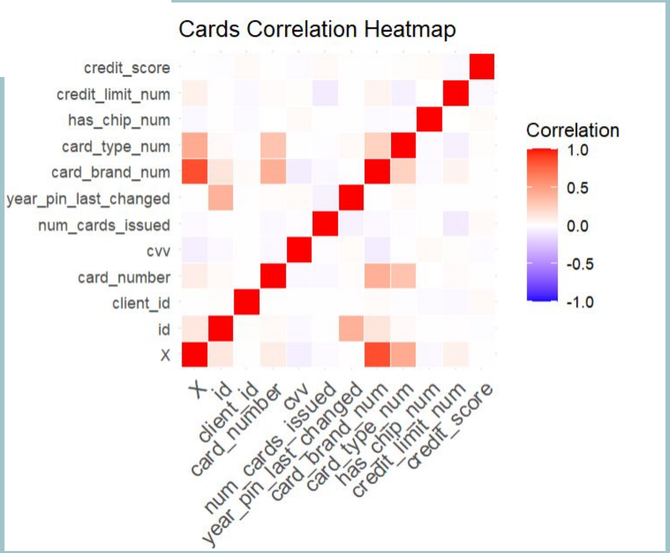
Exploratory
Data
Analysis

Average Credit Score by Total Debt



Visualization Development





Users_data
The correlation matrix heatmap for the Users data set showed retirement_age and num_credit_cards as variables with possible positive correlation. And total_debt was a variable with a possible negative correlation.

Cards_data
The correlation matrix heatmap for the Cards data set showed a small possible positive correlation between the variable num_cards_issued and credit score.

Multiple Regression Model

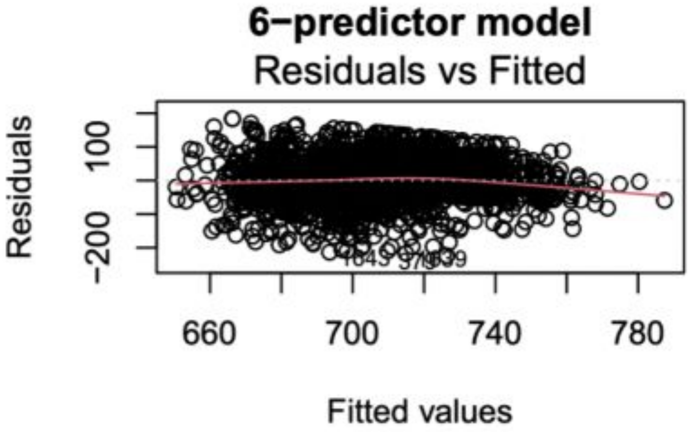

```
##
## Call:
## lm(formula = credit_score ~ current_age + retirement_age + per_capita_income +
##     yearly_income + total_debt + num_credit_cards, data = users_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -217.928  -36.282   -0.069   41.138  187.235
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        5.561e+02  2.688e+01  20.689  < 2e-16 ***
## current_age       -6.107e-01  9.682e-02  -6.307 3.48e-10 ***
## retirement_age     2.258e+00  4.014e-01   5.625 2.12e-08 ***
## per_capita_income -3.420e-04  5.149e-04  -0.664    0.507
## yearly_income      3.263e-04  2.605e-04   1.253    0.210
## total_debt        -1.601e-04  3.465e-05  -4.622 4.05e-06 ***
## num_credit_cards   1.138e+01  1.018e+00  11.175  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 63.83 on 1993 degrees of freedom
## Multiple R-squared:  0.1012, Adjusted R-squared:  0.0985
## F-statistic:  37.4 on 6 and 1993 DF,  p-value: < 2.2e-16
```

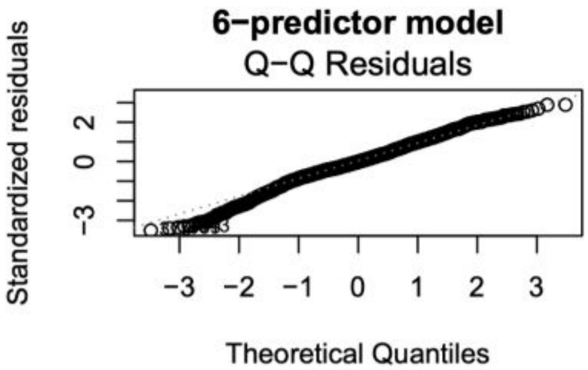
Highly Significant Predictors
($p < 0.001$):

- ➔ current_age (negative effect)
- ➔ retirement_age (positive effect)
- ➔ total_debt (negative effect)
- ➔ num_credit_cards (positive effect)

R-squared = 0.1012
About **10.12%** of the variance in credit score is explained by this model

Model Assumptions





Residuals vs. Fitted Plot

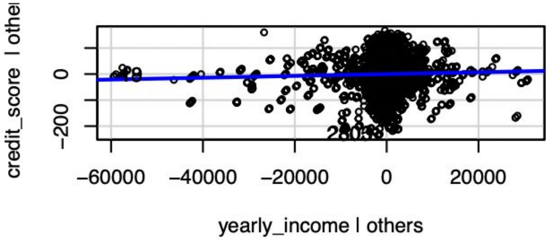
- ❏ Random scatter
- ❏ Constant variance
- ❏ No pattern!

Q-Q Plot of Residuals

- ❏ Diagonal straight line
- ❏ Normality of Residuals

Predictors: current age, retirement age, birth year, yearly income, total debt, number of credit cards

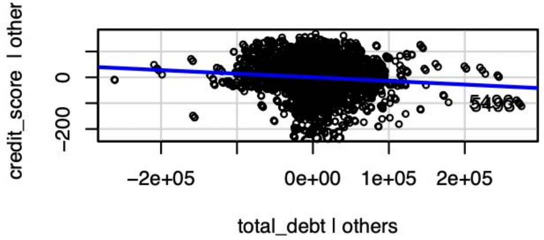
Predictors and Multicollinearity



Yearly Income vs. Credit Score

VIF: 12.855461

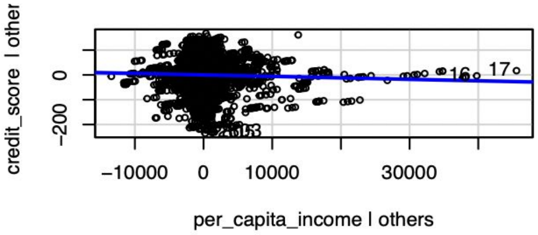
Serious multicollinearity!



Total Debt vs. Credit Score

VIF: 1.547472

Acceptable, little correlation with other variables

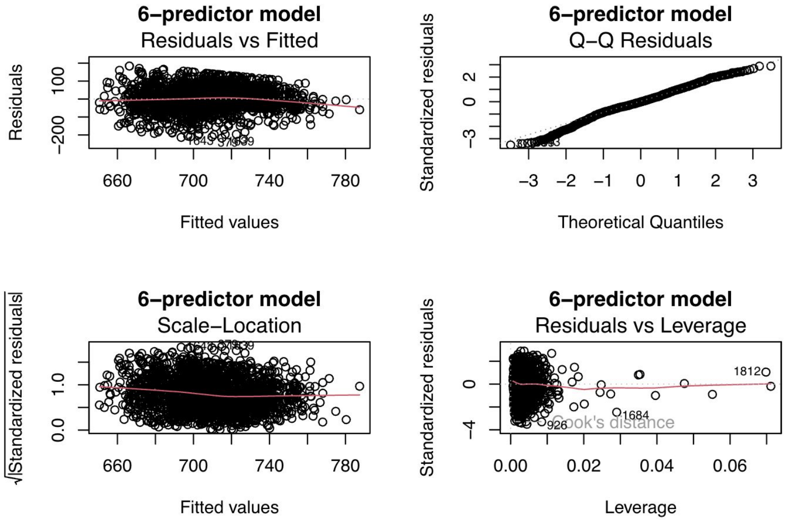


Per Capita Income vs. Credit Score

VIF: 12.564242

Serious multicollinearity!

Model Selection



Method 1: Best Subset Selection

- Evaluate all possible subsets of predictors
- Based on Adjusted R^2, Cp and RSS
    - **Adjusted R^2**: 0.1001
    - **RMSE**: 63.66
- **Conclusions**: the six predictors are moderate predictors of credit score and had the best performance across all metrics

Method 2: Forward Stepwise Regression

- Start with no predictors
- Add predictor that most improves model (lowest AIC)
- Continue until no additional variable improves the model

Method 3: Backwards Stepwise Regression

- Start with all predictors
- Remove predictor that hurts the model the least
- Stop when removing more variables makes the model worse

Limitations & Improvements

- Multicollinearity resulted in a lengthy process of model selection
- Even after model selection, final R^2 value was low
- Investigating further diagnostics such as Cook's distance, leverage points

Thank You