

Abstract

The goal of this project is to build a predictive model using linear regression technique to find Comprehensive Ranking System (CRS) score in Canada's Express Entry system. The CRS score is the cut-off points needed for any individuals to apply for permanent residency in Canada. These scores play vital roles in many international student's lives and are influenced by various factors. I have not included any external factors for this analysis like immigration policy, political changes, and so on. The scope of this project is limited to predicting CRS score solely by the historical express entry draws. I have developed this for the assignment purpose and does not guarantee any accuracy on future express entry draws which involve lot more external factors.

This project demonstrates how linear regression is applied to **predict CRS scores** based on the **three predictor variables like Date, Round type, and number of invitations**. The linear regression model is diagnosed with all the possible interactions and the final model is chosen based on the highest accuracy and low mean squared error. Further improvements can be made to this model by incorporating the external factors and considering other variables and utilizing non-linear modeling techniques.

I have used RSelenium to scrap data from the Express Entry website. Although using RSelenium is out of the scope of this project, I have tried it out of my own interest.

1. Introduction

1.1 Background

The Express entry system in Canada is an immigration pathway that helps candidates to legally immigrate into Canada. The Express entry system is a points-based system called as Comprehensive Ranking System (CRS), every express entry draws conducted by immigration department will have minimum CRS score to be eligible for the immigration. This CRS score is determined based on many factors such as candidate's age, education, work-experience, language skills and so on. On every express entry draws, candidates having CRS scores more than the minimum cut-off will be invited to apply for permanent residency.

1.2 Problem Statement

The CRS cut-off score for every express entry draws fluctuates heavily depending on many factors internally and externally. This project is limited to internal factors like Number of invitations issued, Type of round, time of year influencing the CRS score. In spite of having access to historical data, the model is limited to only internal factors and developing the reliable prediction model including both internal and external factors would be of more benefit.

1.3 Objectives

The objective of this project is to formulate linear regression model to predict the CRS score of the Express entry system of Canada. Using the internal factors and based on the historical express entry rounds, the linear regression model is trained in such a way that can predict CRS score for the future rounds given a few input variables.

1.4 Definition

Linear regression is a technique used in Data analysis tasks which can help predict unknown values by utilizing the known values related to the unknown values. In our project we are predicting

CRS score which is the response variable (i.e) unknown variable by using the various predictor variables (i.e) Known variables such as Round_type, Date, and Number of invitations issued.

The linear regression models hold true based on the assumptions below,

- **Linear relationship:** There should be a linear relationship between independent and dependent variables.
- **Normality:** The residuals should be normalized or should follow the normal distribution.
- **Homoscedasticity:** There should be constant variance or standard deviation from the mean for any value of x.
- **Residuals:** The residuals should be independent of each other.

2. Methods:

2.1 Data Gathering

The Data gathering step is one of the crucial steps in this project as I was involved in web scrapping to download data from the website. I have used one of the interesting packages called RSelenium to scrap web content. As the table I was looking for which contained the information about previous express entry rounds was like a dynamic content which are downloaded by the external URL using JSON, which made me to choose RSelenium over RVEST to do the web scrapping.

As there were 13 pages in total for the target table I was looking for, I used a FOR loop to iterate through every page to download the content. I stored the data from every page into a variable called *Raw_data* as a list and converted it to the data frame finally.

I have written the content I downloaded to csv file and attached along with this project. The code I used for the web scrapping is saved separately in the R file with the name, *Web_Scrapping_RSelenium*.

2.2 Initial Modeling

2.2.1 Data cleaning and Manipulation

I have cleaned the data column wise to make it analytical ready for the linear regression model. I

have used Data wrangling technique on Date column using Lubridate package in R to extract the date from the raw format.

The variable *Round_type* is the categorical variable and therefore I converted that into a factor. The *Round_type* column is now a factor with 12 levels. Below are levels,

Agriculture and agri-food occupations	Canadian Experience Class
Federal skilled Trades	Federal Skilled Worker
French language Trades	General
Healthcare occupations	No Program Specified
Provincial Nominee Program	STEM occupations
Trade occupations	Transport occupations

During the web scrapping process, I encountered the problem in which the data downloaded from the table of different pages didn't have the same datatypes. Therefore, I converted all the data to a character format while web scrapping. Finally, I used mutate function in R to convert the data to its desired datatypes.

The Month and Year columns on the data are very crucial as it can have seasonal patterns in predicting CRS_score. Therefore, I used Year as continuous variable as it can capture any overall trend and can let me know whether the CRS score has linear or non-linear trend over the years. For the month column, as it can strongly have patterns like monthly targets or quarterly immigration targets, it is recommended to consider it as factor variable to consider these patterns for predicting CRS score.

Below is the overall structure of the data that's going to be used for linear regression analysis,

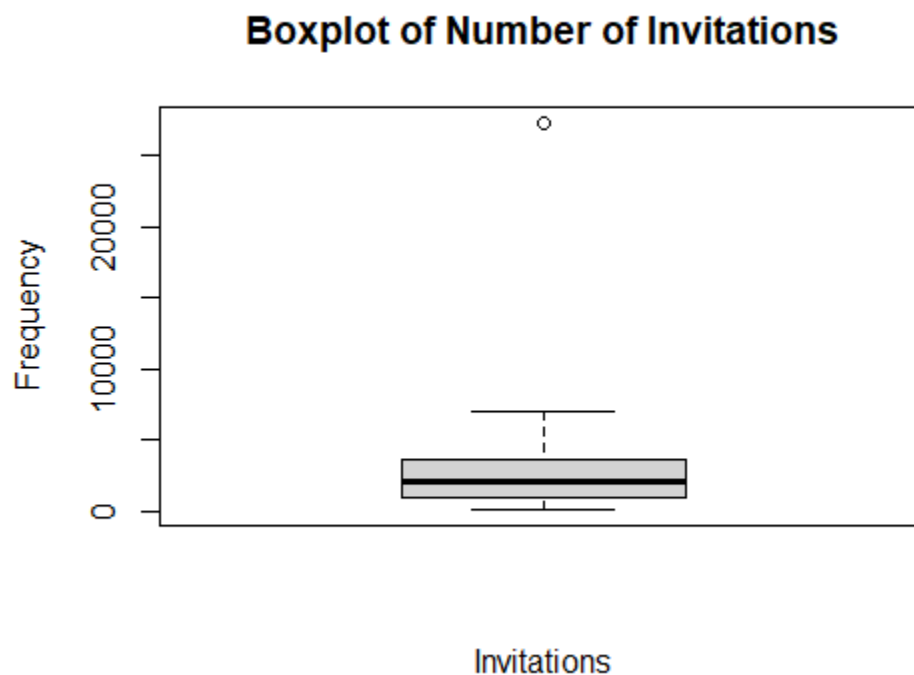
Variable	Datatypes
Month	Factor with 12 levels
Year	numerical values
Round_type	Factor with 12 levels
No_of_invitations	Integer values
CRS_score	Integer values

2.3. Diagnostics

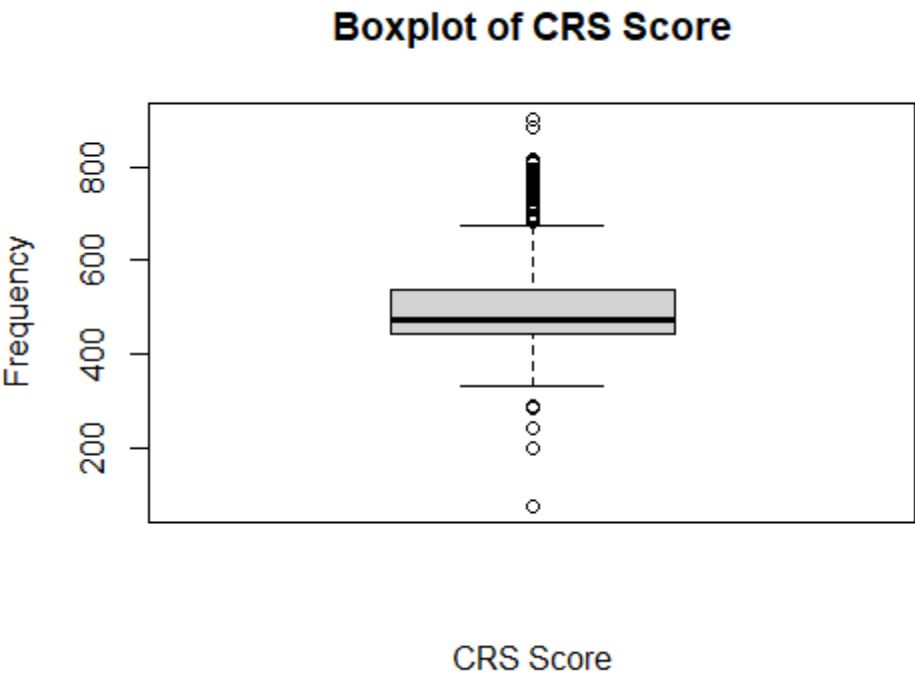
2.3.1 Evaluating the relationship of variables

By evaluating the summary of the table, the No_of_invitations has a huge maximum value which looks strange. Therefore, I will take a look it visually on a box plot to look out for any outliers for all the variables,

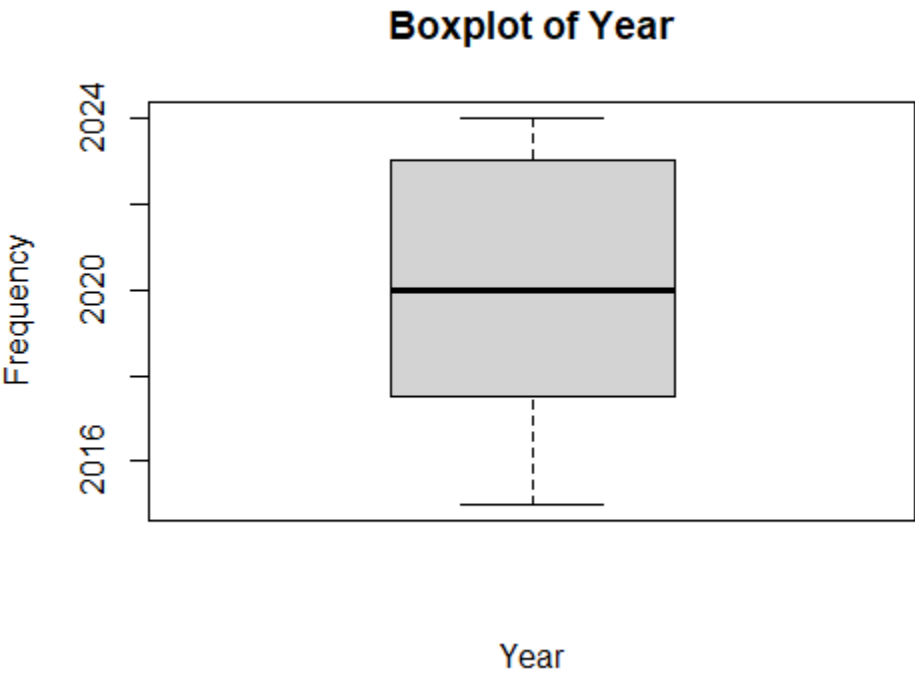
Box_plot for No_of_invitations:



Box_plot for CRS_Score:



Box_plot for Year:



Based on the above box plots, I found two instances for the outliers,

- **No_of_invitations** – The number of invitations has one point estimate which has greater than 27,000 value which is different from rest of the points.
- **CRS_score** – I observed many points for CRS scores are found more 75 percentile of the distribution. But with further analysis it is concluded that the CRS points more than 600 pertains to Provincial Nominee Program, (i.e) it is category specific. Therefore it makes sense to have is larger than 600.

2.3.2 Removing the outliers by IQR method

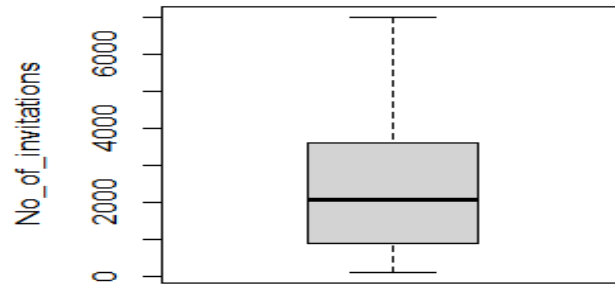
IQR method of removing outliers is one of the prominent techniques used in data analysis to remove any sufficient outliers found in the data. It calculates the first and third quartile which is 25th and 75th percentiles of the data. It further subtracts quartile 3 and quartile 1 to find the inter quartile range.

$$\diamond \text{ IQR} = \text{Q3} - \text{Q1}$$

Now the outliers will be found from the points that fall above $\text{Q3} + 1.5\text{IQR}$ or below $\text{Q1} - 1.5\text{IQR}$.

I have formulated the above technique on *No_of_invitations* column to remove the outliers. Below is the box plot after removing the outliers,

Boxplot of No_of_invitations (Cleaned)

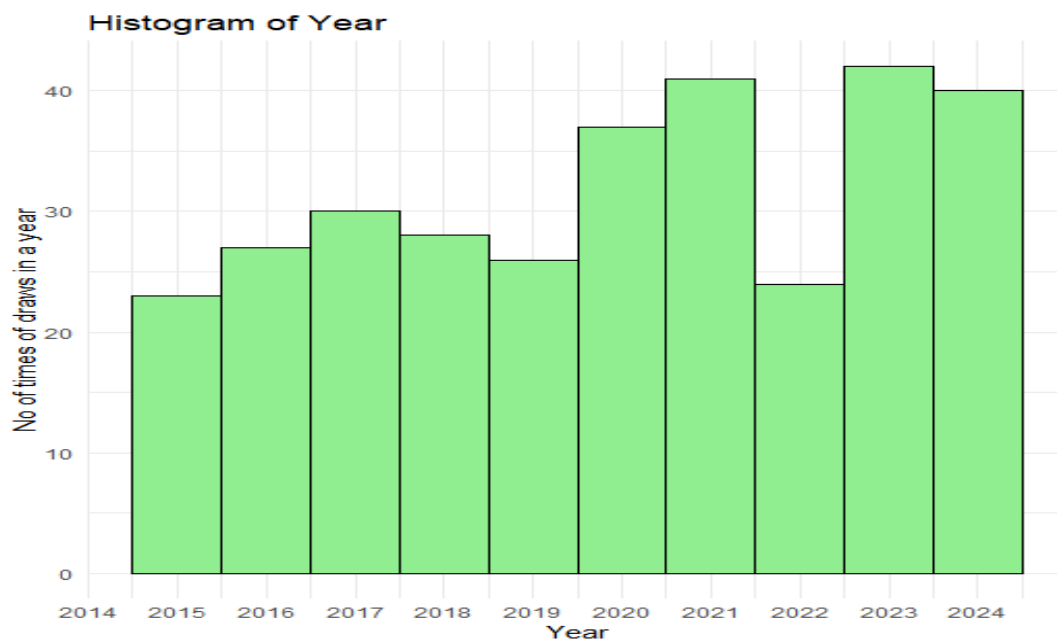


```
> summary(data_cleaned)
```

Month	Year	Round_type	No_of_invitations	CRS_score	
7	: 33	Min. :2015	No Program Specified :167	Min. : 118.0	Min. :199.0
8	: 31	1st Qu.:2017	Provincial Nominee Program : 66	1st Qu.: 897.5	1st Qu.:444.0
4	: 30	Median :2020	Canadian Experience Class : 34	Median :2087.5	Median :471.0
5	: 30	Mean :2020	French language proficiency: 15	Mean :2405.7	Mean :520.7
3	: 28	3rd Qu.:2023	General : 11	3rd Qu.:3600.0	3rd Qu.:538.8
9	: 28	Max. :2024	Federal Skilled Trades : 7	Max. :7000.0	Max. :902.0
(Other)	:138		(Other) : 18		

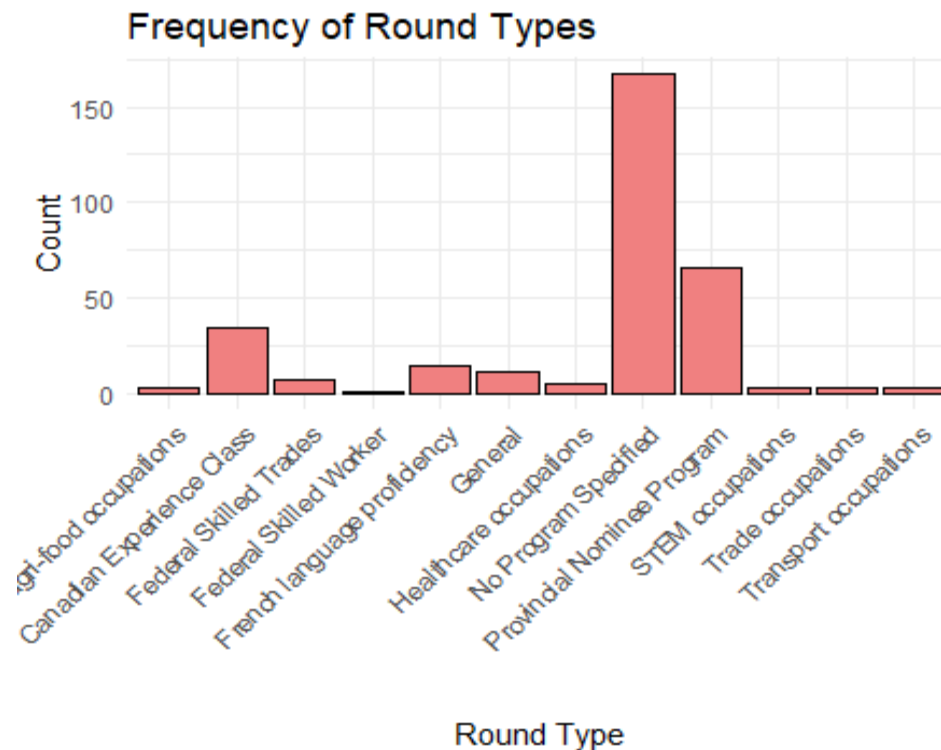
Now the summary and box plot for No_of_invitations looks very good as there are no more outliers.

Histogram for Year:



Based on the above plot it looks like in the year 2023 there were many express entry draws conducted. Nevertheless, as we still in 2024, there are good chances that 2024 will witness higher number of draws than 2023.

Bar graph for Round_types:



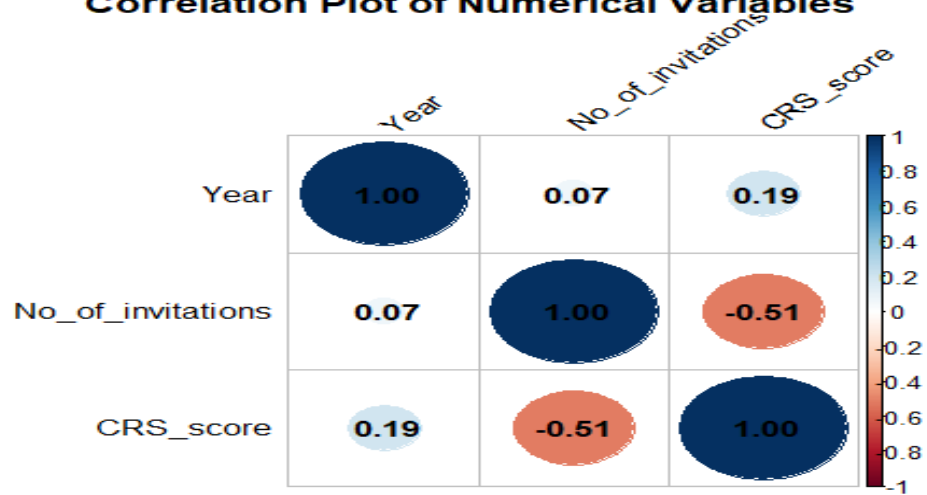
Based on the above bar graph, it is evident that No program specified has more frequency than any draw types.

2.3.3 Multicollinearity check:

The important condition on the linear regression is that there should not be any relationship between predictor variables which is called Multicollinearity. Multicollinearity can highly distort the results of linear regression. Therefore, checking for multicollinearity is a necessary step before formulating the linear regression.

The multicollinearity can be checked by plotting the correlation plot for the independent variable. I have used corrpilot package in R to find the correlation.

Correlation Plot of Numerical Variables

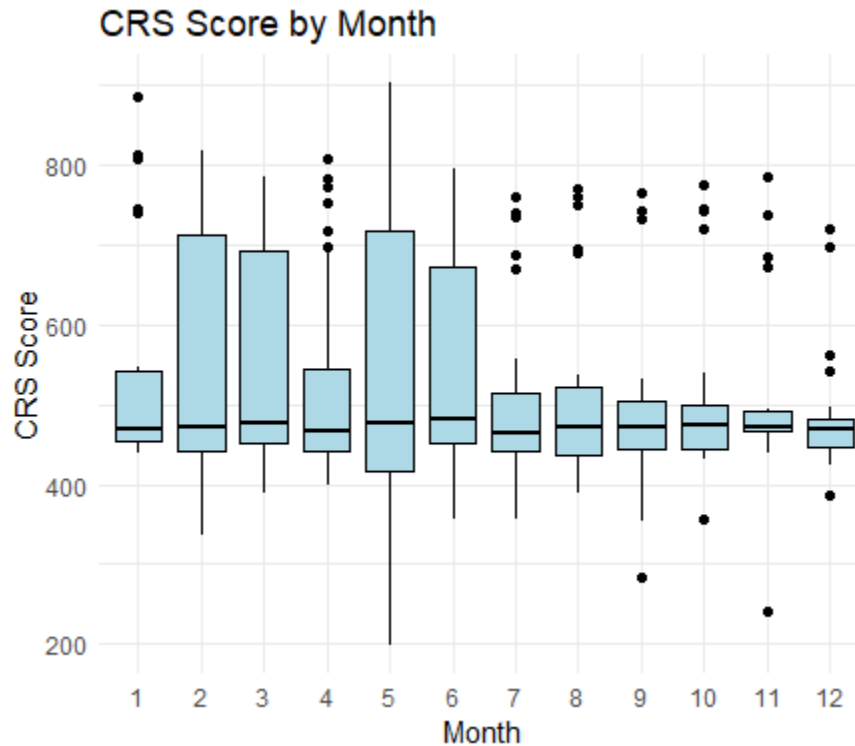


The above plot illustrates that correlation between independent variables. Below are the inferences drawn,

- **Year vs. No_of_invitations:** $r = 0.06$ -> It shows a very weak relationship which implies that changes in year have little to no relationship with No_of_invitations.
- **Year vs CRS_score:** $r = 0.19$ -> It is a weak relationship but holding somewhat positive relationship between. It suggests the CRS score is slightly increasing by time.
- **No_of_invitations vs. CRS_score:** $r = -0.51$ -> It is a negative correlation which shows that CRS score tends to decrease by increase in No_of_invitations.

The above relationship shows that we don't have any significant relationship between predictor variables and therefore multicollinearity doesn't exist between them. Moreover, there is a strong negative correlation between No_of_invitations and CRS score which indicates that it is a significant predictor.

CRS_Score by month:

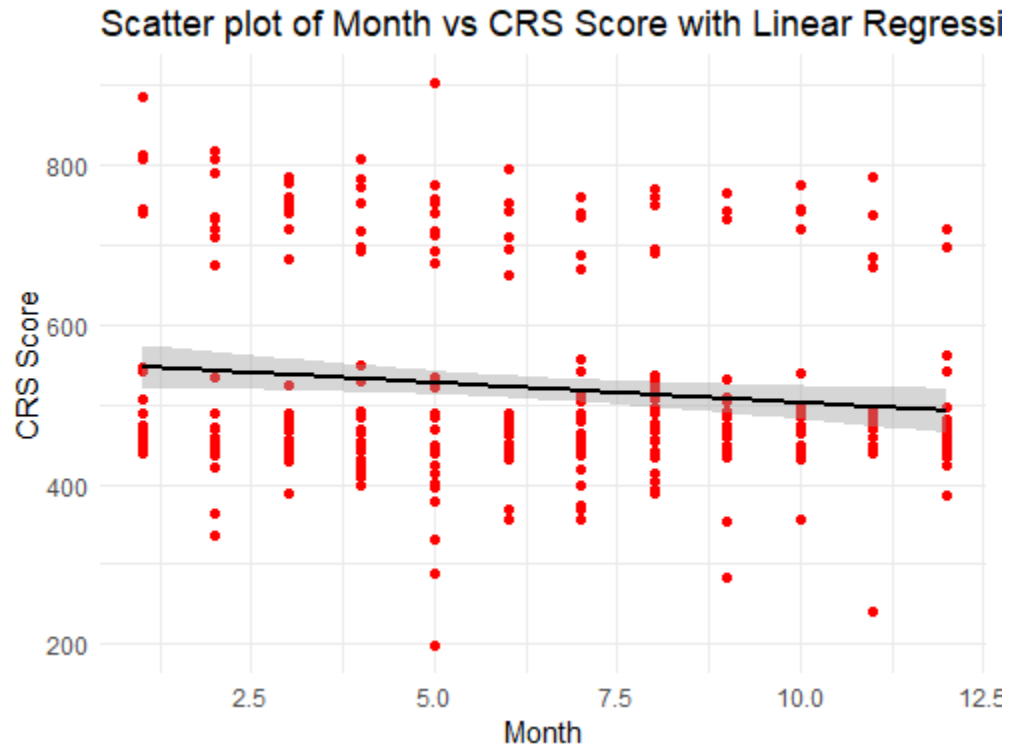


The above plot clearly indicates that there are a lot of outliers every month when the points are above 600, but as we discussed previously it's a PNP category draw which will always be more than 600. Additionally, it is evident that Month 5 is distributed widely indicating that scores are spread and higher chances for the candidates getting picked up.

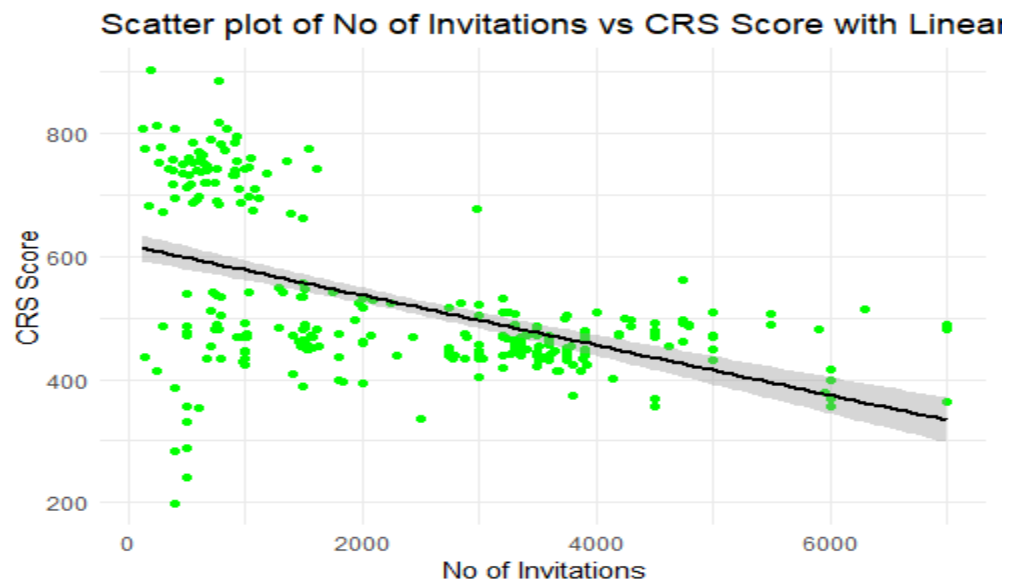
2.3.4 Scatter plot against dependent variable (CRS_score)

It is highly necessary to formulate scatter plot against CRS_score to check whether there is any linear relationship between dependent and independent variables. It is crucial if we are using linear regression models because if there is not any linear relationship then the model may have distorted results. Below are scatter plots for the independent variables against CRS_score,

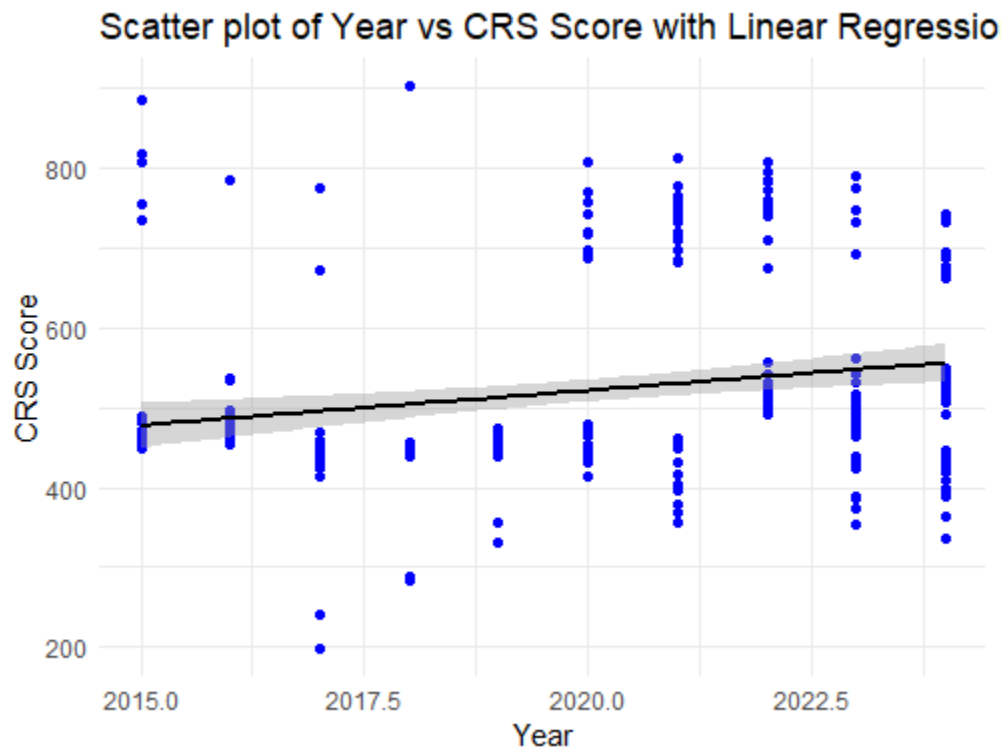
Scatter plot for Month against CRS_score:



Scatter plot for No_of_invitations against CRS_score:



Scatter plot for Year against CRS_score:



Based on all the scatter plots, it is evident that none of the independent variables have a linear relationship against dependent variables. But the reason for non-linearity is due to the presence of the category-based draw, Provincial Nominee Program which will have the minimum score of 600. There are three ways to overcome this tendency,

- i. Assume the PNP draws as outlier and remove it from the dataset as it will distort the entire linear regression results.
- ii. Divide the dataset into two groups and formulate linear regression separately.
- iii. Consider adding interaction terms in the model which can lessen the effects on the regression results.

Although we have three ways to overcome this problem, at this point it is making real sense to include the interaction terms in the regression equation which can potentially minimize the effect of non-linearity.

2.4 Model Selection

It is a standard procedure in the machine learning space to divide data into test data and training data. I have used the ratio of 70% of the data for the training and 30% for the testing.

The splitting of the data will help us in effectively train the models and simultaneously use the same models to predict with the test data and compare the models based on its testing efficiency based on the factors like RMSE, R_squared, and MAE.

To proceed for the linear regression equation, we have formulated five different linear regression models to choose one effective and optimal model for the prediction. Below are the regression models,

1. Simple model
2. Small model
3. Large model
4. Interaction model
5. Polynomial model
6. Subset model

2.4.1 Simple Model

I have formulated the simple model just by having one variable, No_of_invitations to predict the CRS_score. I have used the train data for the model which was splitted earlier. As we saw earlier, both the variables have negative linear relationship between each other and the No_of_invitations variable should have decent power to predict the CRS_score.

Here is the summary for the simple model,

```
Call:
lm(formula = CRS_score ~ No_of_invitations, data = train_data)

Residuals:
    Min       1Q   Median       3Q      Max
-403.91  -64.91  -20.69   87.24  297.95

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    618.59459    13.15727   47.015  <2e-16 ***
No_of_invitations -0.03921     0.00435   -9.014  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 110.7 on 221 degrees of freedom
Multiple R-squared:  0.2688,    Adjusted R-squared:  0.2655
F-statistic: 81.25 on 1 and 221 DF,  p-value: < 2.2e-16
```

The model summary will have the intercept for the independent variables and other factors like Multiple R-squared, Adjusted R-squared, F-stat, df, p-value, and RSE. We will be paying attention to R-squared which tells us the extent the independent variables can explain the variance in the dependent variable. It basically explains how well the model fits the data. As per the above summary, the simple model just has 26.88% of the variation of the dependent variable that can be explained.

2.4.2 Small Model

In the small model, I have included three independent variables such as *No_of_invitations*, *Round_type*, and *Month*. Adding one or more variables to the linear regression equation makes it multiple linear regression.

The addition of the variables will improve the R squared for the model and can predict even better than simple model. Below is the summary of the small model,

```
Call:
lm(formula = CRS_score ~ No_of_invitations + Round_type + Month,
    data = train_data)

Residuals:
    Min       1Q   Median       3Q      Max
-91.59 -27.62  -6.73   17.83  359.52

Coefficients:
(Intercept)                410.940539  Std. Error 33.521461  t value 12.259  Pr(>|t|) < 2e-16 ***
No_of_invitations          -0.014446    Std. Error  0.002876   t value -5.024  1.12e-06 ***
Round_typeCanadian Experience Class 135.736743    Std. Error 34.746037   t value  3.907  0.000128 ***
Round_typeFederal Skilled Trades   -98.224837    Std. Error 40.915619   t value -2.401  0.017281 *
Round_typeFrench language proficiency  53.589721    Std. Error 36.150663   t value  1.482  0.139808
Round_typeGeneral                172.622155    Std. Error 36.385984   t value  4.744  3.98e-06 ***
Round_typeHealthcare occupations  109.786400    Std. Error 49.120565   t value  2.235  0.026521 *
Round_typeNo Program Specified    126.791454    Std. Error 32.441373   t value  3.908  0.000127 ***
Round_typeProvincial Nominee Program 363.203580    Std. Error 32.185554   t value 11.285  < 2e-16 ***
Round_typeSTEM occupations        160.035439    Std. Error 44.475111   t value  3.598  0.000404 ***
Round_typeTrade occupations        52.769934    Std. Error 43.601098   t value  1.210  0.227596
Round_typeTransport occupations    55.356350    Std. Error 42.955841   t value  1.289  0.198998
Month2                          -0.824934    Std. Error 17.417374   t value -0.047  0.962271
Month3                         -22.296794    Std. Error 16.341427   t value -1.364  0.173964
Month4                         -39.407997    Std. Error 16.660143   t value -2.365  0.018967 *
Month5                         -32.030308    Std. Error 16.156383   t value -1.983  0.048790 *
Month6                         -34.666057    Std. Error 17.019715   t value -2.037  0.042986 *
Month7                         -38.895258    Std. Error 16.599088   t value -2.343  0.020100 *
Month8                         -21.956939    Std. Error 16.790462   t value -1.308  0.192475
Month9                         -19.223247    Std. Error 19.386371   t value -0.992  0.322598
Month10                        -13.575636    Std. Error 18.643034   t value -0.728  0.467350
Month11                        -28.539441    Std. Error 20.538039   t value -1.390  0.166199
Month12                        -19.160210    Std. Error 18.646999   t value -1.028  0.305416
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 52.15 on 200 degrees of freedom
Multiple R-squared:  0.8531,    Adjusted R-squared:  0.837
F-statistic: 52.8 on 22 and 200 DF,  p-value: < 2.2e-16
```

As discussed earlier, addition of the variable drastically increased the efficiency of the model with 85.31% of R squared.

2.4.3 Large Model

I have formulated the large model with the same as small model and adding another new independent variable “Year” additionally. Below is the summary for the large model,

```
lm(formula = CRS_score ~ No_of_invitations + Year + Month + Round_type,
    data = train_data)

Residuals:
    Min       1Q   Median       3Q      Max
-95.42 -24.90  -5.26   15.80  368.45

Coefficients:
(Intercept)                -1.207e+04  Std. Error  4.268e+03  t value -2.827  0.00518 **
No_of_invitations          -2.001e-02    Std. Error  3.404e-03   t value -5.878  1.73e-08 ***
Year                        6.166e+00    Std. Error  2.109e+00   t value  2.923  0.00386 **
Month2                      3.376e+00    Std. Error  1.716e+01   t value  0.197  0.84420
Month3                      -2.015e+01    Std. Error  1.606e+01   t value -1.255  0.21105
Month4                      -3.591e+01    Std. Error  1.640e+01   t value -2.190  0.02968 *
Month5                      -2.841e+01    Std. Error  1.591e+01   t value -1.786  0.07568 .
Month6                      -3.372e+01    Std. Error  1.671e+01   t value -2.018  0.04497 *
Month7                      -4.147e+01    Std. Error  1.632e+01   t value -2.542  0.01180 *
Month8                      -2.779e+01    Std. Error  1.660e+01   t value -1.674  0.09576 .
Month9                      -1.942e+01    Std. Error  1.903e+01   t value -1.021  0.30871
Month10                     -1.481e+01    Std. Error  1.831e+01   t value -0.809  0.41951
Month11                     -2.066e+01    Std. Error  2.034e+01   t value -1.016  0.31096
Month12                     -1.266e+01    Std. Error  1.844e+01   t value -0.687  0.49312
Round_typeCanadian Experience Class  1.720e+02    Std. Error  3.629e+01   t value  4.739  4.09e-06 ***
Round_typeFederal Skilled Trades   -6.292e+01    Std. Error  4.194e+01   t value -1.500  0.13515
Round_typeFrench language proficiency  6.438e+01    Std. Error  3.568e+01   t value  1.804  0.07270 .
Round_typeGeneral                1.785e+02    Std. Error  3.577e+01   t value  4.990  1.32e-06 ***
Round_typeHealthcare occupations  1.196e+02    Std. Error  4.834e+01   t value  2.474  0.01420 *
Round_typeNo Program Specified    1.750e+02    Std. Error  3.586e+01   t value  4.879  2.18e-06 ***
Round_typeProvincial Nominee Program 3.805e+02    Std. Error  3.214e+01   t value 11.837  < 2e-16 ***
Round_typeSTEM occupations        1.791e+02    Std. Error  4.415e+01   t value  4.058  7.11e-05 ***
Round_typeTrade occupations        6.275e+01    Std. Error  4.294e+01   t value  1.461  0.14550
Round_typeTransport occupations    5.881e+01    Std. Error  4.218e+01   t value  1.394  0.16482
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

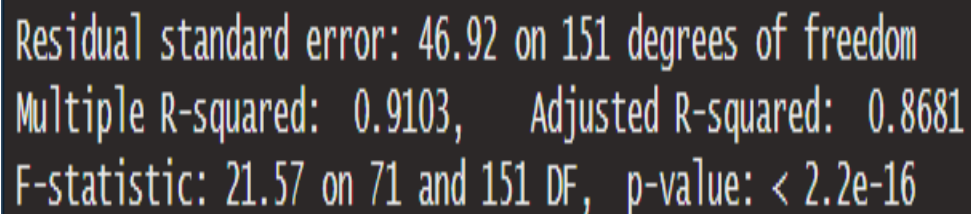
Residual standard error: 51.2 on 199 degrees of freedom
Multiple R-squared:  0.8592,    Adjusted R-squared:  0.8429
F-statistic: 52.78 on 23 and 199 DF,  p-value: < 2.2e-16
```


The new variable added, “Year” is the continuous numerical variable and therefore the addition didn’t improve the model much. The Large model has 85.92% R squared, which is the better model until now.

2.4.4 Interaction model

It’s obvious that the relationship between dependent and independent variables are not so linear because the PNP category draw’s minimum CRS, which is more than 600. Therefore, adding an interaction by including the variable “Round_type” to all other independent variable can improve the model performance.

Adding the “Round_type” to the regression equation can help the model to consider the categories available in the express entry system and can predict the CRS scores accordingly. Below is the summary of the interaction model,

A screenshot of a regression model summary with a dark background and light-colored text. The text displays the residual standard error, multiple and adjusted R-squared values, and the F-statistic with its degrees of freedom and p-value.

```
Residual standard error: 46.92 on 151 degrees of freedom  
Multiple R-squared: 0.9103, Adjusted R-squared: 0.8681  
F-statistic: 21.57 on 71 and 151 DF, p-value: < 2.2e-16
```

The interaction model has the highest R squared value which is of 91.03% and thereby explaining most of the variation found in the dependent variable.

2.4.5 Polynomial model

With reference to the scatter plot for the No_of_invitations, it was found evident that there was not a clear linear relationship between the variables. Although we concluded that it is due to the presence of the category draws, I took a different approach of including the polynomial into the equation. Below is the summary of the model performance,

```
Residual standard error: 47.63 on 188 degrees of freedom
Multiple R-squared:  0.8848,    Adjusted R-squared:  0.864
F-statistic: 42.49 on 34 and 188 DF,  p-value: < 2.2e-16
```

Adding the polynomial degree to the No_of_invitations doesn't help much as the model's R squared decreased to 88.48%. Therefore, it concludes that adding different variables, polynomial, or an interaction cannot always improve the performance of the model.

2.4.6 Subset model (K-fold cross validation)

The subset model is a linear regression model to predict CRS score based on all the variables, but the model is trained using K-fold to improve the performance of the model. I have used 10-fold to improve the model prediction. Below is the summary of the model,

```
> print(subset_model)
Linear Regression

223 samples
 4 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 201, 199, 201, 201, 201, 202, ...
Resampling results:

    RMSE      Rsquared   MAE
51.40681  0.8267984  33.90645

Tuning parameter 'intercept' was held constant at a value of TRUE
```

The R-squared has again decreased to 82.67%, and has not improved much after the K-fold cross validation.

2.4.7 Model Evaluation

To compare the models effectively, I have calculated RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error) and R-squared for every model to know better on how the models are predicting the errors. I have created the metrics for each model and combined it on single table to show all the model's performance,

Model	RMSE	R_squared	MAE
Simple Model	108.64	0.27	87.14
Small Model	54.68	0.85	33.97
Large Model	54.05	0.86	30.76
Interaction Model	58.42	0.91	34.09
Polynomial + Interaction Model	61.92	0.88	34.63
Subset model	51.41	0.83	NA

The simple model performs the worst as it has the highest Root mean squared error of 108.6 and mean absolute error of 87.14. and the lowest R-squared of around 26%. With a much greater R-squared of 85%, and MAE of 33.97, and an RMSE of 54.68, the Small Model shows a substantial improvement.

With an R-squared of 86%, and MAE of 30.76, and an RMSE of 54.05, the Large Model outperforms the Small Model but by a small margin. The R-squared of 91%, the Interaction Model's RMSE (58.42) and MAE (34.09) are higher than those of the big and small models. Following with an R-squared of 88%, the Polynomial + Interaction Model has a higher RMSE of 61.92 and MAE of 34.63. Lastly, the Subset Model's R-squared 83% is lower than the big model. When evaluating the prediction and accuracy, both Interaction model and Polynomial model have better results compared to other models.

Overall, I would conclude the selection process by picking Interaction model as it can predict the CRS_score better by identifying the variation in the depended variable and taking interaction of Round_type into consideration.

3. Prediction and summary

An interaction model was used to predict a scenario where the following variables were set:

Number of invitations: 4,500

Round type: “STEM Occupations”

Year: 2024

Month: December

Based on these inputs, the model predicted a CRS score of 470. This predicted score aligns well with the current trends, suggesting that the model is reasonably accurate for real-time scenario predictions. The close alignment with known data provides confidence in the model's ability to generalize and predict CRS scores in similar future scenarios.

4. Conclusion

For this project, we created and assessed a number of linear regression models that used factors like No_of_invitations, Round_type, Year, and Month to predict the CRS score in the Canadian immigration system. With the largest error rates and the lowest R-squared value, the evaluation showed that the Simple Model did the poorest. On the other hand, because they could take into consideration the interactions between variables, the Interaction Model and Polynomial model showed better predictive power, with R-squared values of 91% and 88%, respectively. When it

came to projecting a CRS score of 470 for a scenario including 4,500 invites for STEM professions in December 2024, the Interaction Model ultimately proved to be the most successful model.

Reference:

Kavita. (2024, September 30). Linear Regression: A Comprehensive guide. Analytics Vidhya.

<https://www.analyticsvidhya.com/blog/2021/10/everything-you-need-to-know-about-linear-regression/>

Linear regression. (n.d.). <http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm>

Statistics Solutions. (2024, April 19). Linear Regression Analysis: Exploring correlation and directionality - Statistics Solutions. <https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/how-to-conduct-linear-regression/>